Understanding tuberculosis drug resistance, disease phenotype and transmission by mycobacterial genome analysis

Carolien Ruesen

Understanding tuberculosis drug resistance, disease phenotype and transmission by mycobacterial genome analysis

Carolien Ruesen

The work presented in this thesis was carried out within the Radboud Institute for Health Sciences.

ISBN
978-94-028-1715-7

Design/lay-out
Promotie In Zicht, Arnhem

Print
Ipskamp Printing, Enschede

Understanding tuberculosis drug resistance, disease phenotype and transmission by mycobacterial genome analysis

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op
dinsdag 5 november 2019
om 10:30 uur precies

door

Carolina Johanna Ruesen
geboren op 26 juli 1988
te Wehl

Promotoren
Prof. dr. R. van Crevel
Prof. dr. M. A. Huynen

Copromotoren
Dr. L. Chaidir (Universitas Padjadjaran, Bandung, Indonesië)
Dr. J. van Ingen

Manuscriptcommissie
Prof. dr. M. van den Heuvel
Prof. dr. W. Bitter (Vrije Universiteit Amsterdam)
Dr. J.L.A. Hautvast

Contents

1

General introduction

Tuberculosis

Tuberculosis is the number one cause of death in humans due to a single infectious agent[1], with an estimated 10 million new cases and 1.6 million deaths in 2017[2]. Tuberculosis is caused by infection with *Mycobacterium tuberculosis*, which is spread from a person with the respiratory form of the disease to another person via air droplets generated by coughing and sneezing. After exposure to *M. tuberculosis*, the outcome can vary (**Figure 1.1**). Following infection, the bacteria can be cleared through the action of the innate immune system (early clearance); the infection can rapidly progress to active tuberculosis disease, or it can be contained in a latent form with no clinical signs of disease[2]. One fourth of the world's population has this latent form of tuberculosis infection, which may or may not reactivate up to several decades after initial exposure[3]. In addition, not only the outcome of *M. tuberculosis* infection is variable, the course of active disease can also vary. Pulmonary tuberculosis and various forms of extrapulmonary disease such as tuberculous meningitis and miliary tuberculosis are all presentations of active tuberculosis disease.

Various factors contribute to the heterogeneity of *M. tuberculosis* infection and tuberculosis disease. For example, weakened immunity, such as in human immunodeficiency virus (HIV)-positive individuals, is a strong risk factor for rapid progression from *M. tuberculosis* infection to active disease[4]. To date, most of the research efforts on tuberculosis have focused on host and environmental factors influencing transmission, drug resistance, and susceptibility to infection and disease[5]. However, it is becoming increasingly clear that bacterial factors are also important. Not only is the genetic diversity of *M. tuberculosis* higher than previously thought[6], knowledge is also expanding that this genomic diversity translates into relevant disease phenotypic variation. Over the past decades, studies have shown that *M. tuberculosis* genotype families show differences in terms of their geographical spread[7,8], immune responses in the patient[9], and virulence[6,10]. Virulence in tuberculosis relates to the ability of the bacteria to survive in the face of the host immune responses, their capacity to cause lung damage, to survive in aerosols outside the host, and lastly to successfully transmit and infect a new host[11,12]. Altogether, this suggests that *M. tuberculosis* genetic variation may play an important role in the success of the pathogen, which has been coexisting with mankind since hunter-gatherer times for an estimated 70,000 years[13].

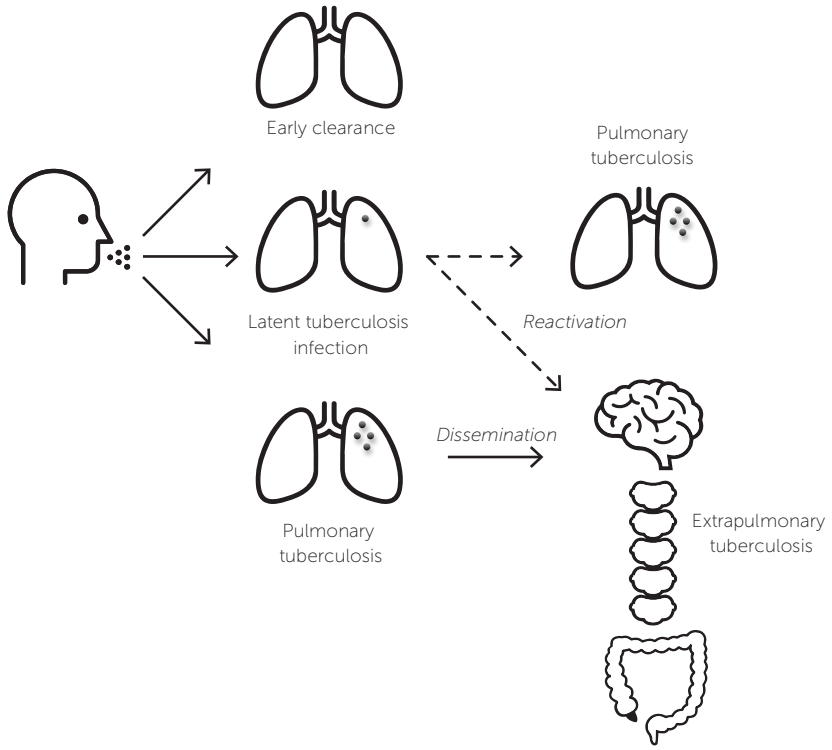


Figure 1.1. The different outcomes and disease manifestations after exposure to *M. tuberculosis*

Human tuberculosis is mainly caused by *M. tuberculosis* and *M. africanum*. Both are human-adapted members of a bigger *Mycobacterium tuberculosis* complex (MTBC), which emerged as a professional pathogen from an environmental Mycobacterium through stepwise adaptation to an intracellular milieu. MTBC has little sequence variation compared with other bacteria and some of the genotyping tools applied to other bacterial pathogens are uninformative in MTBC[14]. Hence, classifying MTBC clinical strains requires analytical methods that capture genomic diversity in a more comprehensive manner. Whole genome sequencing, in contrast with other typing techniques, allows indexing of MTBC diversity across the whole genome, capturing as much variation as possible and enabling detailed analyses of the evolutionary forces driving this diversity.

Today, based on whole genome sequence analyses, we know that the human-adapted members of MTBC can be further divided into seven phylogenetic lineages: Lineage 1 to 4 and Lineage 7 belong to *M. tuberculosis sensu stricto*, whereas Linage 5 and 6 are traditionally known as *M. africanum* West Africa 1 and 2, respectively[15]. These lineages can be grouped into evolutionarily "ancient" and "modern" lineages, with a deletion in the TbD1 genomic region found exclusively in the evolutionarily "modern" lineages, serving as a genetic marker separating the two groups. The geographical spread of these lineages differs markedly, with the modern lineages (Lineage 2, 3 and 4) exhibiting a global distribution and the ancient lineages (Lineage 1, 5, 6 and 7) a strong geographical restriction[6]. Lineage 2 (also known as the East-Asian lineage) and Lineage 4 (Euro-American lineage) are the most widespread globally. Comparisons with *M. canetti* and other smooth tubercle bacilli, which are phylogenetically the closest relatives of MTBC[16], have shown that many individual genomic features including point mutations in key genes, as well as epistatic interactions between these features might have contributed to the evolution of *M. tuberculosis* towards an obligate human pathogen[16]. This in contrast to many other bacterial pathogens in which pathogenicity results from the presence or absence of specific "virulence factors".

Strains of MTBC differ genetically in their content of single nucleotide polymorphisms (SNPs), small insertions and deletions, mobile and repetitive elements, large genomic deletions, and large duplications. Whole genome sequencing can reveal all sources of genomic variation. The complete genome sequence of the best-characterized strain of *M. tuberculosis*, H37Rv (belonging to Lineage 4), has been determined and published for the first time by Cole and colleagues in 1998[17]. The *M. tuberculosis* H37Rv genome comprises approximately 4.4 million base pairs and contains around 4,000 genes. *M. tuberculosis* differs from other bacteria in that a very large proportion of its coding potential accounts for the production of enzymes involved in lipogenesis and lipolysis, and for two families of glycine-rich proteins with characteristic Pro-Glu (PE) or Pro-Pro-Glu (PPE) motifs that may represent a source of antigenic variation[17,18]. Based on whole genome sequencing analyses of all MTBC lineages, we know that on average, two human-adapted MTBC strains differ by about 1,200 SNPs (0.03% of the genome when excluding repetitive elements)[6]. Strains from the same lineage show less genetic diversity than strains from different lineages, with the least diversity for strains belonging to Lineage 7 (on average 230 SNPs difference between any two strains belonging to this lineage), and largest for strains belonging to Lineage 1 (730 SNPs on average)[6].

In addition to exploring and quantifying the genomic diversity in MTBC to get a better understanding of the evolution and molecular epidemiology, whole genome sequencing data can be used together with computational methods to predict the impact of this

variation on a range of clinical or immunological phenotypes. Ultimately, differences among MTBC strains are determined by their genomic differences. In this thesis I have used whole genome sequencing data to link *M. tuberculosis* genetic variation to three relevant phenotypes: drug resistance (**Chapter 2, 3** and **4**), disease localisation (**Chapter 5**) and transmission (**Chapter 6**), which I will discuss in the following sections.

Antituberculosis drug resistance

Drug-resistant *M. tuberculosis* was first documented in the late 1940s, shortly after the introduction of the first anti tuberculosis drug, streptomycin[19]. However, drug-resistant tuberculosis did not receive global attention until the 1990s when large outbreaks of multidrug-resistant tuberculosis were identified in the United States and Europe[20-22]. It has since become clear that drug-resistant tuberculosis, caused by rifampicin-resistant, multidrug-resistant or extensively drug-resistant strains (**Box 1.1**), is a growing concern and threatens global tuberculosis control. Treatment for patients with multidrug-resistant or extensively drug-resistant tuberculosis is prolonged and expensive, and associated with morbidity and mortality[23,24]. The drugs used are toxic and poorly tolerated, and adverse events are common and may be severe and irreversible[25,26].

Patients can carry a drug-resistant *M. tuberculosis* strain as a result of two processes: acquisition or transmission of resistance. Noncompliance or misuse of antituberculosis drugs, such as monotherapy or the addition of single drugs to failing regimens, can result in the emergence of resistant mutants (acquired resistance). Transmission of such resistant strains to another person may result in infection and eventually disease (primary or transmitted resistance)[24]. Outbreaks of highly fatal drug-resistant tuberculosis have been documented in several settings, especially those in which the prevalence of HIV infection is high[32-34].

Phenotypic testing is the standard for drug susceptibility testing despite its challenges, which are felt most acutely in resource-limited settings. These challenges include the slow growth rate of *M. tuberculosis*, the high-level biosafety infrastructure required and uncertainties around the proposed clinical breakpoints for some drugs[35]. The mechanism underlying drug resistance in *M. tuberculosis* differs from other bacterial species that often acquire resistance through horizontal gene transfer systems such as plasmid exchange. In *M. tuberculosis*, drug resistance is solely conferred by de novo mutations in genes coding for drug targets or for proteins involved in drug metabolic pathways, possibly preceded by efflux pump up-regulation. Changes in the DNA of *M. tuberculosis* usually result from point mutations, indels, or, more rarely, large deletions[36]. Together with *M. tuberculosis*' low mutation rate of an estimated 0.3-0.5

Box 1.1. Definitions of drug-resistant tuberculosis

Drug-susceptible tuberculosis
Tuberculosis caused by *M. tuberculosis* strains that are susceptible to all first-line anti-tuberculosis drugs. The first-line standard regimen that is currently recommended for the treatment of drug-susceptible tuberculosis is based on a 2-month intensive phase with four drugs (isoniazid, rifampicin, pyrazinamide and ethambutol) followed by a 4-month consolidation phase with two drugs (isoniazid and rifampicin)[27].

Rifampicin-resistant tuberculosis
Tuberculosis caused by *M. tuberculosis* strains that are resistant to rifampicin.

Multidrug-resistant tuberculosis
Tuberculosis caused *by M. tuberculosis* strains that are resistant to at least isoniazid and rifampicin, the two most effective first-line antituberculosis drugs. Treatment is more costly and toxic and has lower cure rates than drug-susceptible tuberculosis[28,29], although this has improved since the introduction of bedaquiline[30].

Extensively drug-resistant tuberculosis
Tuberculosis caused by multidrug-resistant *M. tuberculosis* strains that are additionally resistant to any fluoroquinolone and any second-line injectable drug (i.e., kanamycin, amikacin, or capreomycin). Cure and survival rates are worse for extensively drug-resistant tuberculosis than for multidrug-resistant tuberculosis[31].

substitutions per genome per year[36], this makes whole genome sequencing a well-suited technique to study *M. tuberculosis*, creating a unique opportunity for genotypic drug susceptibility testing. Early results from an extensive study linking whole genome sequencing data to phenotypic drug susceptibility testing have provided information for a pilot of drug-susceptibility testing based on whole genome sequencing in the United Kingdom[37].

Tuberculosis disease phenotype

The result of exposure to *M. tuberculosis* depends on the interplay between the innate and adaptive immune mechanisms of the human host on the one hand, and the bacterial mechanisms to evade or antagonize these immune responses on the other hand. Early clearance can be defined as the eradication of infecting *M. tuberculosis* by the innate immune system before an adaptive immune response develops, and was

recently found to occur in 25% of exposed individuals[38]. A persistent infection (latent tuberculosis) develops in those who fail to clear the infection, with an approximately 10% lifetime risk of progression to active disease, highest in the first 18 months after infection[39]. Although active tuberculosis predominantly affects the lungs (pulmonary tuberculosis), it can cause disease in any other organ (extrapulmonary disease). The classic clinical features of pulmonary tuberculosis include cough, sputum production, haemoptysis, loss of appetite, weight loss, fever, and night sweats[40]. The wide variation in clinical manifestations, disease severity and outcome is one of the most intriguing aspects of tuberculosis and has been mainly attributed to host factors, among others HIV infection, diabetes, smoking, age, and malnutrition. However, host factors do not fully explain the observed variation and evidence is compiling that *M. tuberculosis* strain diversity is also important[5,41].

M. tuberculosis strain differences have been characterized *in vitro* in macrophage infection models and *in vivo* in animals[42] and differences have been defined by bacterial growth in cells or organs, the death of infected cells or animals, and differences in the histopathology of infected animal tissues[43]. Highly virulent *M. tuberculosis* isolates appeared to grow faster[44], to cause more lung damage and mortality[43,45], and were more capable to transmit[46] than attenuated or low virulence strains. Differences between strains in clinical settings have been demonstrated with respect to their inflammatory response, the severity of the disease, transmission rate and disease presentation[8]. In addition, results from clinical and epidemiological studies suggested that strains from Lineage 5 and Lineage 6 are metabolically different, grow slower, and are less virulent than the other human-adapted MTBC lineages. Similarly, at least certain groups of Lineage 2 and Lineage 4 strains have been shown to be more virulent in terms of disease severity and human-to-human transmission, and strains belonging to the East-Asian/Beijing lineage were more likely to progress to active tuberculosis disease and were associated with extra-pulmonary tuberculosis, multidrug resistance, treatment failure, and relapse[47-49].

These findings support that, next to human and environmental factors, *M. tuberculosis* strain diversity contributes to the variable outcome of infection and disease in human tuberculosis. However, the specific *M. tuberculosis* genetic factors determining the variation in disease phenotypes remain largely unknown, and differences between *M. tuberculosis* lineages do not explain the full spectrum of phenotypic variation observed. Whole genome sequencing offers the possibility to study *M. tuberculosis* genomic diversity at the highest possible resolution, at the level of individual nucleotide changes, therewith also capturing differences within *M. tuberculosis* lineages[50]. A better understanding of the genetic determinants of disease phenotypes, what I aim for in this thesis, would in turn aid in the understanding of tuberculosis' pathogenesis.

Mycobacterium tuberculosis transmission

Transmission from one person to another is crucial for *M. tuberculosis*' survival. In order to be able to transmit, the bacterium needs to cause disease, leading to destruction of lung tissue and consequent expulsion of aerosolized *M. tuberculosis* bacilli. Disease development is a function of the host's immunocompetence; individuals with HIV, for example, are at increased risk of progression to active disease. However, disease development is also a reflection of the evolutionary strategy of *M. tuberculosis* as a pathogen, which during human existence has needed to ensure transmission to the next host. *M. tuberculosis* has to undertake a delicate balancing act: cause enough disease to ensure transmission but not so much that patients rapidly die, taking the pathogen's progeny with them[27].

Preventing tuberculosis transmission is one of the spear points of tuberculosis control, but has proven to be extremely difficult. One of the reasons is the absence of an effective vaccine against tuberculosis. The bacillus Calmette-Guérin (BCG) vaccine has been widely used and is effective at preventing severe childhood forms of tuberculosis, but protection wanes by adolescence and the protective efficacy in adults is highly variable. The elimination of tuberculosis cannot reasonably be achieved by treatment of individual patients; hence an improved vaccine is required. The main challenge in developing an improved vaccine against tuberculosis has been the lack of understanding of correlates of a protective immune response. So far, vaccine development in tuberculosis has focused on inducing protective adaptive immunity to prevent progression to active tuberculosis disease, but the observation that some individuals do not develop *M. tuberculosis* infection even after heavy exposure to tuberculosis, suggests that the innate immune system is important for preventing initial infection with *M. tuberculosis*. Recent evidence suggests that BCG vaccination also plays a role in the natural protection or early clearance of *M. tuberculosis*. Alternative approaches aimed at boosting innate immunity could improve vaccine development to protect against *M. tuberculosis* infection, so contributing to tuberculosis prevention.

Transmission of tuberculosis is currently measured by genotyping of the isolates, epidemiological contact investigation, or a combination of the two[51]. However, epidemiological data are often challenging to obtain, and genotyping data are difficult to interpret without them. Whole genome sequencing facilitates identifying likely transmission events among patients without the requirement of epidemiological data; based on the principle that patients infected with genetically similar strains are considered to be part of the same transmission cluster[52,53]. Using this 'DNA fingerprinting' technique, many studies to date have investigated the association between *M. tuberculosis* genetic variation and transmission. However, two processes are involved in transmission;

initial infection of a contact, followed by progression of latent tuberculosis infection to active disease, and these cannot be distinguished by studying genetic clustering. This distinction is important when examining early clearance, which relates to direct transmission: passing infection from index case to contact persons[54]. *M. tuberculosis* could promote transmission by evading early clearance in the contacts, and the extent to which this evasion is successful is possibly genetically determined. Understanding which *M. tuberculosis* strain variation is associated with a more, or less successful innate immune response in the host will aid the identification of new vaccine targets and is topic of one of the studies in this thesis.

Aim and scope of the thesis

This thesis aims at better understanding of the *M. tuberculosis* genetic factors determining drug resistance, disease presentation and transmission of tuberculosis. We examined the genomes of *M. tuberculosis* strains isolated from more than 1,000 patients from Indonesia, Peru and Romania, and linked genomic data with clinical metadata. The thesis consists of two parts. The first part focuses on antituberculosis drug resistance, combining *M. tuberculosis* genetic data with phenotypic drug susceptibility testing results, as well as information on patient-related risk factors for drug resistance. The second part addresses the effect of *M. tuberculosis* genetic variation on tuberculosis disease presentation and on transmission within households (**Figure 1.2**).

Drug resistance

The first part of this thesis describes the use of *M. tuberculosis* whole genome sequencing data to predict antituberculosis drug resistance in high-burden, low-resource countries. In Indonesia, which has the third-highest tuberculosis burden in the world, diagnosis of drug-resistant tuberculosis is difficult as *M. tuberculosis* culture is not routinely performed and phenotypic drug susceptibility testing is only available in reference laboratories. In **Chapter 2**, we examined for the first time the resistance-conferring mutations to first- and second-line drugs in Indonesia using whole genome sequencing, and assessed the level of agreement between genotypic and phenotypic drug susceptibility testing.

Although whole genome sequencing can identify mutations associated with antituberculosis drug resistance, the impact of the many resistance mutations on the minimum inhibitory concentration remains unclear[55]. The level of resistance that a single nucleotide polymorphism causes, reflected by the minimum inhibitory concentration, is important for clinicians treating patients in order to determine whether to increase the dosage or change the regimen. Therefore, in **Chapter 3** we examined the genomes

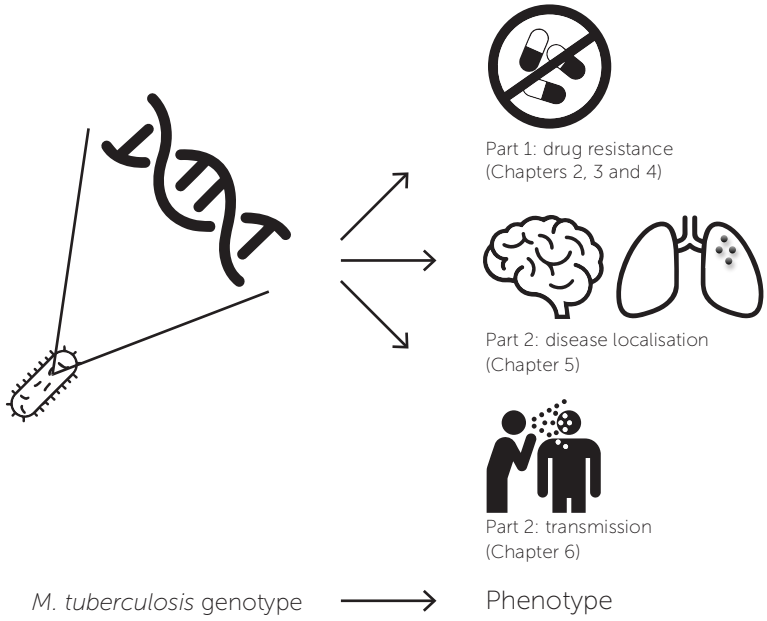


Figure 1.2. The *M. tuberculosis* genotype was linked to multiple tuberculosis phenotypes in my thesis

of phenotypically drug-resistant *M. tuberculosis* isolates from Romanian patients for drug resistance mutations, and linked these to the measured minimum inhibitory concentrations.

Phenotypic antituberculosis drug resistance has been associated with diabetes mellitus comorbidity[56] and this may contribute to the poor treatment outcomes among tuberculosis patients with diabetes, but a basic understanding of the molecular mechanism behind this association is lacking. As opposed to phenotypic drug susceptibility testing, whole genome sequencing has the potential to unlock valuable information on the specific mutations underlying resistance, transmission clusters and on the phylogenetic background of drug-resistant *M. tuberculosis* strains. In **Chapter 4**, we used whole genome sequencing to assess genotypic drug resistance in *M. tuberculosis* isolates from Indonesian and Peruvian patients with and without diabetes. We examined if drug resistance mutations were more frequently found in isolates from patients with diabetes, compared to those without, after adjusting for other risk factors for drug resistance.

Tuberculosis disease phenotype and transmission

Many studies have attempted to link host or environmental factors to various tuberculosis disease phenotypes, but together they do not fully explain the observed spectrum. In this part of the thesis we studied the effect of bacterial factors, using whole genome sequencing. Animal studies have shown that *M. tuberculosis* strains differ in their ability to disseminate to and survive in the central nervous system[57,58], and *in vitro* studies have identified *M. tuberculosis* genes that were crucial for invading an artificial blood-brain barrier[59]. Much less is known about the role of *M. tuberculosis* genotype on the disease phenotype in humans. In **Chapter 5,** we therefore examined the effect of the infecting *M. tuberculosis* genotype on the susceptibility to tuberculous meningitis, compared to pulmonary tuberculosis. We used a novel approach of counting homoplasy events to detect genetic variants associated with either the pulmonary or meningeal tuberculosis presentation.

In addition to disease phenotype, *M. tuberculosis* genetic variation could also influence the risk of *M. tuberculosis* transmission. *M. tuberculosis* factors can increase the risk of transmission in two ways: first; through the index case, by inducing a disease phenotype that favours transmission. A higher bacillary load with more immunopathology and worse cavitary disease results in increased spread of *M. tuberculosis*[41]. Second; through the contact, by resistance to the contact's innate immune system[60]. Understanding these pathogen factors is essential to understanding on the one hand the virulence of *M. tuberculosis* in the index case, related to outgrowth and cavities, and on the other hand the virulence of *M. tuberculosis* in the contact, related to resistance to innate immunity. However, epidemiological evidence for *M. tuberculosis* genotype influencing risk of infection is scarce. Therefore, in **Chapter 6**, we investigated whether *M. tuberculosis* genotype influences the disease phenotype in the contact, or the success with which it can escape the innate immune defence of the contact, using a household-based case-contact study in Bandung, Indonesia.

Lastly, in **Chapter 7**, the findings of this thesis are summarized and discussed, and future challenges and opportunities of whole genome sequencing of *M. tuberculosis* are delineated.

References

1. Paulson T. Epidemiology: A mortal foe. Nature. 2013;502(7470):S2-3.
2. Global Tuberculosis Report 2018. Geneva: World Health Organization.
3. Houben RM, Dodd PJ. The Global Burden of Latent Tuberculosis Infection: A Re-estimation Using Mathematical Modelling. PLoS Med. 2016;13(10):e1002152.
4. Kwan CK, Ernst JD. HIV and tuberculosis: a deadly human syndemic. Clin Microbiol Rev. 2011;24(2):351-76.
5. Comas I, Gagneux S. The past and future of tuberculosis research. PLoS Pathog. 2009;5(10):e1000600.
6. Coscolla M, Gagneux S. Consequences of genomic diversity in *Mycobacterium tuberculosis*. Semin Immunol. 2014;26(6):431-44.
7. Gagneux S, Small PM. Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. The Lancet Infectious Diseases. 2007;7(5):328-37.
8. Parwati I, van Crevel R, van Soolingen D. Possible underlying mechanisms for successful emergence of the *Mycobacterium tuberculosis* Beijing genotype strains. Lancet Infect Dis. 2010;10(2):103-11.
9. van Laarhoven A, Mandemakers JJ, Kleinnijenhuis J, Enaimi M, Lachmandas E, Joosten LA, et al. Low induction of proinflammatory cytokines parallels evolutionary success of modern strains within the *Mycobacterium tuberculosis* Beijing genotype. Infect Immun. 2013;81(10):3750-6.
10. Coscolla M, Gagneux S. Does *M. tuberculosis* genomic diversity explain disease diversity? Drug Discov Today Dis Mech. 2010;7(1):e43-e59.
11. Orgeur M, Brosch R. Evolution of virulence in the *Mycobacterium tuberculosis* complex. Curr Opin Microbiol. 2018;41:68-75.
12. Brites D, Gagneux S. The Nature and Evolution of Genomic Diversity in the *Mycobacterium tuberculosis* Complex. Adv Exp Med Biol. 2017;1019:1-26.
13. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, et al. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. Nat Genet. 2013;45(10):1176-82.
14. Comas I, Homolka S, Niemann S, Gagneux S. Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. PLoS One. 2009;4(11):e7815.
15. Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, Narayanan S, et al. Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. Proc Natl Acad Sci U S A. 2006;103(8):2869-73.
16. Supply P, Marceau M, Mangenot S, Roche D, Rouanet C, Khanna V, et al. Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of *Mycobacterium tuberculosis*. Nat Genet. 2013;45(2):172-9.
17. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature. 1998;393(6685):537-44.
18. Brennan MJ. The PE multigene family- a 'molecular mantra' for mycobacteria. Trends in Microbiology. 2002.
19. Crofton J, Mitchison DA. Streptomycin resistance in pulmonary tuberculosis. Br Med J. 1948;2(4588):1009-15.
20. Monno L, Carbonara S, Costa D, Angarano G, Coppola S, Quarto M, et al. Emergence of drug-resistant *Mycobacterium tuberculosis* in HIV-infected patients. The Lancet. 1991;337(8745):852.
21. CDC. Nosocomial Transmission of Multidrug-Resistant Tuberculosis Among HIV-Infected Persons—Florida and New York, 1988-1991. JAMA: The Journal of the American Medical Association. 1991;266(11):1483.
22. Frieden TR, Sterling T, Pablos-Mendez A, Kilburn JO, Cauthen GM, Dooley SW. The emergence of drug-resistant tuberculosis in New York City. N Engl J Med. 1993;328(8):521-6.
23. Zürcher K, Ballif M, Fenner L, Borrell S, Keller PM, Gnokoro J, et al. Drug susceptibility testing and mortality in patients treated for tuberculosis in high-burden countries: a multicentre cohort study. The Lancet Infectious Diseases. 2019;19(3):298-307.
24. Dheda K, Gumbo T, Maartens G, Dooley KE, McNerney R, Murray M, et al. The epidemiology, pathogenesis, transmission, diagnosis, and management of multidrug-resistant, extensively drug-resistant, and incurable tuberculosis. The Lancet Respiratory Medicine. 2017;5(4):291-360.
25. Shean K, Streicher E, Pieterson E, Symons G, van Zyl Smit R, Theron G, et al. Drug-associated adverse events and their relationship with outcomes in patients receiving treatment for extensively drug-resistant tuberculosis in South Africa. PLoS One. 2013;8(5):e63057.

26. Wu S, Zhang Y, Sun F, Chen M, Zhou L, Wang N, et al. Adverse Events Associated With the Treatment of Multidrug-Resistant Tuberculosis: A Systematic Review and Meta-analysis. Am J Ther. 2016;23(2):e521-30.
27. Dheda K, Barry CE, Maartens G. Tuberculosis. The Lancet. 2015.
28. Bastos ML, Lan Z, Menzies D. An updated systematic review and meta-analysis for treatment of multi-drug-resistant tuberculosis. Eur Respir J. 2017;49(3).
29. Migliori GB, Dheda K, Centis R, Mwaba P, Bates M, O'Grady J, et al. Review of multidrug-resistant and extensively drug-resistant TB: global perspectives with a focus on sub-Saharan Africa. Trop Med Int Health. 2010;15(9):1052-66.
30. Mbuagbaw L, Guglielmetti L, Hewison C, Bakare N, Bastard M, Caumes E, et al. Outcomes of Bedaquiline Treatment in Patients with Multidrug-Resistant Tuberculosis. Emerg Infect Dis. 2019;25(5):936-43.
31. Kvasnovsky CL, Cegielski JP, Erasmus R, Siwisa NO, Thomas K, der Walt ML. Extensively drug-resistant TB in Eastern Cape, South Africa: high mortality in HIV-negative and HIV-positive patients. J Acquir Immune Defic Syndr. 2011;57(2):146-52.
32. Cohen KA, Abeel T, Manson McGuire A, Desjardins CA, Munsamy V, Shea TP, et al. Evolution of Extensively Drug-Resistant Tuberculosis over Four Decades: Whole Genome Sequencing and Dating Analysis of *Mycobacterium tuberculosis* Isolates from KwaZulu-Natal. PLoS Med. 2015;12(9):e1001880.
33. Walker TM, Merker M, Knoblauch AM, Helbling P, Schoch OD, van der Werf MJ, et al. A cluster of multi-drug-resistant *Mycobacterium tuberculosis* among patients arriving in Europe from the Horn of Africa: a molecular epidemiological study. The Lancet Infectious Diseases. 2018;18(4):431-40.
34. Khan PY, Yates TA, Osman M, Warren RM, van der Heijden Y, Padayatchi N, et al. Transmission of drug-resistant tuberculosis in HIV-endemic settings. The Lancet Infectious Diseases. 2019;19(3):e77-e88.
35. McNerney R, Zignol M, Clark TG. Use of whole genome sequencing in surveillance of drug resistant tuberculosis. Expert Rev Anti Infect Ther. 2018;16(5):433-42.
36. Eldholm V, Balloux F. Antimicrobial Resistance in *Mycobacterium tuberculosis*: The Odd One Out. Trends Microbiol. 2016;24(8):637-48.
37. Walker TM, Kohl TA, Omar SV, Hedge J, Del Ojo Elias C, Bradley P, et al. Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study. The Lancet Infectious Diseases. 2015;15(10):1193-202.
38. Verrall AJ, Schneider M, Alisjahbana B, Apriani L, van Laarhoven A, Koeken V, et al. Early clearance of *Mycobacterium tuberculosis* is associated with increased innate immune responses. J Infect Dis. 2019.
39. Andrews JR, Noubary F, Walensky RP, Cerda R, Losina E, Horsburgh CR. Risk of progression to active tuberculosis following reinfection with *Mycobacterium tuberculosis*. Clin Infect Dis. 2012;54(6):784-91.
40. Lawn SD, Zumla AI. Tuberculosis. The Lancet. 2011;378(9785):57-72.
41. Coscolla M. Biological and Epidemiological Consequences of MTBC Diversity. Adv Exp Med Biol. 2017;1019:95-116.
42. Prozorov AA, Fedorova IA, Bekker OB, Danilenko VN. The virulence factors of *Mycobacterium tuberculosis*: Genetic control, new conceptions. Russian Journal of Genetics. 2014;50(8):775-97.
43. Dormans J, Burger M, Aguilar D, Hernandez-Pando R, Kremer K, Roholl P, et al. Correlation of virulence, lung pathology, bacterial load and delayed type hypersensitivity responses after infection with different *Mycobacterium tuberculosis* genotypes in a BALB/c mouse model. Clin Exp Immunol. 2004;137(3):460-8.
44. Theus SA, Cave MD, Eisenach KD. Intracellular macrophage growth rates and cytokine profiles of *Mycobacterium tuberculosis* strains with different transmission dynamics. J Infect Dis. 2005;191(3):453-60.
45. Manca C, Tsenova L, Bergtold A, Freeman S, Tovey M, Musser JM, et al. Virulence of a *Mycobacterium tuberculosis* clinical isolate in mice is determined by failure to induce Th1 type immunity and is associated with induction of IFN-alpha /beta. Proc Natl Acad Sci U S A. 2001;98(10):5752-7.
46. Marquina-Castillo B, Garcia-Garcia L, Ponce-de-Leon A, Jimenez-Corona ME, Bobadilla-Del Valle M, Cano-Arellano B, et al. Virulence, immunopathology and transmissibility of selected strains of *Mycobacterium tuberculosis* in a murine model. Immunology. 2009;128(1):123-33.
47. Caws M, Thwaites G, Dunstan S, Hawn TR, Lan NT, Thuong NT, et al. The influence of host and bacterial genotype on the development of disseminated disease with *Mycobacterium tuberculosis*. PLoS Pathog. 2008;4(3):e1000034.

48. Thwaites G, Caws M, Chau TT, D'Sa A, Lan NT, Huyen MN, et al. Relationship between *Mycobacterium tuberculosis* genotype and the clinical phenotype of pulmonary and meningeal tuberculosis. J Clin Microbiol. 2008;46(4):1363-8.
49. Parwati I, Alisjahbana B, Apriani L, Soetikno RD, Ottenhoff TH, van der Zanden AG, et al. *Mycobacterium tuberculosis* Beijing genotype is an independent risk factor for tuberculosis treatment failure in Indonesia. J Infect Dis. 2010;201(4):553-7.
50. Stucki D, Brites D, Jeljeli L, Coscolla M, Liu Q, Trauner A, et al. *Mycobacterium tuberculosis* lineage 4 comprises globally distributed and geographically restricted sublineages. Nat Genet. 2016;48(12):1535-43.
51. Yates TA, Khan PY, Knight GM, Taylor JG, McHugh TD, Lipman M, et al. The transmission of *Mycobacterium tuberculosis* in high burden settings. The Lancet Infectious Diseases. 2016;16(2):227-38.
52. Walker TM, Monk P, Smith EG, Peto TE. Contact investigations for outbreaks of *Mycobacterium tuberculosis*: advances through whole genome sequencing. Clin Microbiol Infect. 2013;19(9):796-802.
53. Walker TM, Ip CLC, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. The Lancet Infectious Diseases. 2013;13(2):137-46.
54. Verrall AJ, G. Netea M, Alisjahbana B, Hill PC, van Crevel R. Early clearance of *Mycobacterium tuberculosis*: a new frontier in prevention. Immunology. 2014;141(4):506-13.
55. Walker TM, Merker M, Kohl TA, Crook DW, Niemann S, Peto TE. Whole genome sequencing for M/XDR tuberculosis surveillance and for resistance testing. Clin Microbiol Infect. 2017;23(3):161-6.
56. Tegegne BS, Mengesha MM, Teferra AA, Awoke MA, Habtewold TD. Association between diabetes mellitus and multi-drug-resistant tuberculosis: evidence from a systematic review and meta-analysis. Syst Rev. 2018;7(1):161.
57. Be NA, Lamichhane G, Grosset J, Tyagi S, Cheng QJ, Kim KS, et al. Murine model to study the invasion and survival of *Mycobacterium tuberculosis* in the central nervous system. J Infect Dis. 2008;198(10):1520-8.
58. Be NA, Bishai WR, Jain SK. Role of *Mycobacterium tuberculosis* pknD in the pathogenesis of central nervous system tuberculosis. BMC Microbiol. 2012;12:7.
59. Jain SK, Paul-Satyaseela M, Lamichhane G, Kim KS, Bishai WR. *Mycobacterium tuberculosis* invasion and traversal across an in vitro human blood-brain barrier as a pathogenic mechanism for central nervous system tuberculosis. J Infect Dis. 2006;193(9):1287-95.
60. Simmons JD, Stein CM, Seshadri C, Campo M, Alter G, Fortune S, et al. Immunological mechanisms of human resistance to persistent *Mycobacterium tuberculosis* infection. Nat Rev Immunol. 2018;18(9):575-89.

PART ONE

Antituberculosis drug resistance

2

Use of whole genome sequencing to predict *Mycobacterium tuberculosis* drug resistance in Indonesia

Carolien Ruesen*, Lidya Chaidir*, Bas E. Dutilh, Ahmad R. Ganiem, Anggriani Andryani, Lika Apriani, Martijn A. Huynen, Rovina Ruslami, Philip C. Hill, Reinout van Crevel, Bachti Alisjahbana.

* Equal contribution

Journal of Global Antimicrobial Resistance. 2019; 16:170-177.

Abstract

Objectives

Whole-genome sequencing (WGS) is rarely used for drug resistance testing of *Mycobacterium tuberculosis* in high-endemic settings. Here we present the first study from Indonesia, which has the third highest tuberculosis (TB) burden worldwide, with <50% of drug-resistant cases currently detected.

Methods

WGS was applied for strains from 322 human immunodeficiency virus (HIV)-negative adult TB patients. Phenotypic drug susceptibility testing (DST) was performed for a proportion of the patients.

Results

Using WGS, mutations associated with drug resistance to any TB drug were identified in 51 (15.8%) of the 322 patients, including 42 patients (13.0%) with no prior TB treatment (primary resistance). Eight isolates (2.5%) were multidrug-resistant (MDR) and one was extensively drug-resistant (XDR). Most mutations were found in *katG* (n = 18), *pncA* (n = 18), *rpoB* (n = 10), *fabG1* (n = 9) and *embB* (n = 9). Agreement of WGS-based resistance and phenotypic DST to first-line drugs was high for isoniazid and rifampicin but was lower for ethambutol and streptomycin. Drug resistance was more common in Indo- Oceanic lineage strains (37.5%) compared with Euro-American (18.2%) and East-Asian lineage strains (10.3%) (P = 0.044), but combinations of multiple mutations were most common among East-Asian lineage strains (P = 0.054).

Conclusions

These data support the potential use of WGS for more rapid and comprehensive prediction of drug-resistant TB in Indonesia. Future studies should address potential barriers to implementing WGS, the distribution of specific resistance mutations, and the association of particular mutations with endemic *M. tuberculosis* lineages in Indonesia.

Introduction

Drug-resistant tuberculosis (TB) threatens the global control of TB in many parts of the world. There were an estimated 580 000 cases of multidrug-resistant TB (MDR-TB) in 2015, with <50% being detected and even fewer receiving appropriate treatment[1,2]. Culture-based drug susceptibility testing (DST), the gold standard to diagnose drug-resistant TB, is technically difficult and takes about 4-6 weeks after the isolation of *Mycobacterium tuberculosis*. Furthermore, inconsistencies in DST results are widely reported, especially for ethambutol (EMB) and second-line drugs[3,4], whilst phenotypic DST for pyrazinamide (PZA) requires different protocols[5]. Molecular-based DST testing using the Xpert MTB/ RIF assay can rapidly detect most rifampicin (RIF) resistance[6], but this assay has incomplete sensitivity for RIF, does not examine resistance for other TB drugs and may fail to detect heteroresistance[7,8].

Whole-genome sequencing (WGS) has been shown to be a potential tool for reliable prediction of the drug susceptibility phenotype of *M. tuberc*ulosis isolates within a clinically relevant timeframe[9]. However, so far WGS is mostly applied in well-resourced, low TB burden settings. Because of rapid advances in WGS technology and its decreasing cost and turnaround time, WGS is now becoming accessible in limited-resourced, high TB burden countries[10]. In the current study, WGS was applied to predict drug resistance in a selection of *M. tuberculosis* isolates from Indonesia, which has the third-highest TB burden in the world. In Indonesia, diagnosis of drug-resistant TB is challenging as *M. tuberculosis* culture is not routinely performed and DST is only available in certain reference laboratories. To date, the contribution and concordance of phenotypic and genotypic DST in Indonesia is unknown. This study aimed to describe resistance-conferring mutations to first- and second-line TB drugs in Indonesia using WGS and to examine their concordance with phenotypic DST and their distribution among different *M. tuberculosis* lineages.

Materials and Methods

Selection of *M. tuberculosis* isolates and drug susceptibility testing

WGS was performed on a random sample of archived *M. tuberculosis* isolates from 322 human immunodeficiency virus (HIV)-negative adult patients (216 with pulmonary TB and 106 with meningeal TB) with complete medical information. Pulmonary TB patients had been diagnosed in Hasan Sadikin Hospital (Bandung, Indonesia) between 2012–2013 and in the TB-HIV Research Centre of Universitas Padjadjaran (Bandung, Indonesia) between 2013–2015. TB meningitis patients had been diagnosed at Hasan Sadikin Hospital between 2006–2013[11]. Specimens from each patient were processed

accordingly and were inoculated on solid Ogawa medium or in MODS (Microscopic Observation of Drug Susceptibility) liquid medium[12]. Positive cultures from each method were subcultured and aliquots were archived at -80 °C prior to DNA extraction.

Xpert MTB/RIF was available in the study setting only after 2012 and was accessible only for patients with suspected MDR-TB according to Indonesian national guidelines. Phenotypic DST was not performed routinely for all patients but only when requested by treating physicians. DST was performed in the provincial referral laboratory using the proportion method on Löwenstein–Jensen (LJ) medium at concentrations of 40.0 µg/mL for RIF, 0.2 µg / mL for isoniazid (INH), 2.0 µg /mL for EMB and 4.0 µg /mL for streptomycin (STR). Briefly, a 1.0 McFarland standard isolate suspension was serially diluted 10-fold (from 10^{-1} to 10^{-5}) in sterile distilled water. Dilutions 10^{-3} and 10^{-5} were inoculated, respectively, onto LJ slants with and without drugs, and were incubated at 37 °C. The results were read at 28 days and up to 42 days, depending on the control growth. An isolate was considered resistant to a given drug when growth of ≥1% above the control was observed in drug-containing medium. DST for second-line drugs was not available during patient inclusion for this study. Phenotypic DST was repeated for several isolates that had tested INH-susceptible but that harboured mutations conferring resistance to INH in the *katG*315 gene. In these cases, *M. tuberculosis* was subcultured from frozen isolates onto Ogawa slopes prior to DST. The study protocols for the inclusion of patients and for bioanalysis were approved by the Ethical Committee of the Faculty of Medicine of Universitas Padjadjaran.

Whole genome sequencing and analysis

Frozen isolates were subcultured on Ogawa slopes and mycobacterial DNA was extracted for sequencing using Ultra- Clean1 Microbial DNA Isolation Kit (MO BIO Laboratories, Carlsbad, CA) following the manufacturer's protocol, or using the cetyltrimethylammonium bromide (CTAB) method of DNA purification. The concentration and purity of extracted DNA were measured using a NanoDrop™ 2000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA) and the intactness of DNA was checked by agarose gel electrophoresis.

M. tuberculosis DNA was sequenced on an Illumina HiSeq 2000 instrument (Illumina Inc., San Diego, CA) using 2 x 100-bp paired-end reads. Sequencing was performed at BGI (Hong Kong). After sequencing, the raw FASTQ sequence reads were filtered, including removal of adapter sequences, contamination, and low-quality reads that had more than 10% N base calls or had a quality score ≤4 in more than 40% of the bases. Five TB meningitis strains and four pulmonary TB strains were contaminated, based on a low GC content and were excluded from further analyses. Sequencing coverage was determined using the FastQC v.0.10.1 (http://www.bioinformatics.

babraham.ac.uk/projects/fastqc/) quality control tool. The proportion of bases sequenced with a sequencing error rate of ≤1% per base ranged from 93% to 97% per genome. The average depth of coverage for the remaining 322 sequenced strains was 121.1, and the average percentage of bases covered by at least one read was 98.9%. The sequence reads were aligned to reference strain *M. tuberculosis* H37Rv (GenBank accession no. NC_000962.3) and variants were called using Breseq software v.0.27.1[13]. Mutations with low-quality evidence (i.e. possible mixed read alignment) were not included.

In silico determination of drug resistance

Raw FASTQ sequencing files were uploaded to TB Profiler, an online tool to determine drug resistance *in silico*[10]. It uses raw sequence data as input and compares identified single nucleotide polymorphisms (SNPs) and indels to a curated list of 1,325 drug resistance mutations and displays related output. The precision of TB Profiler's curated mutation catalogue for predicting resistance had been assessed using six geographically distinct data sets from China, Pakistan, Malawi, Portugal, Russia and Canada[10], and its accuracy compared with other *in silico* drug resistance prediction tools has been proven recently[14].

Phylogeny construction

A phylogenetic tree was constructed to determine the evolutionary relationship of the isolates. All 29,199 variable positions identified by breseq across the 322 *M. tuberculosis* sequences were extracted and concatenated into a single alignment. Solely for the purpose of creating the phylogenetic tree, SNPs occurring in PE/PPE genes and genes related to mobile elements (listed in Table S2.1) were excluded to avoid any concern about inaccuracies in the read alignment in these parts of the genome. In addition, SNPs in an additional 40 genes previously associated with drug resistance[10] were also removed to exclude the possibility that homoplasy of drug resistance mutations would significantly affect the phylogeny[15]. After applying these filters to the initial set of 29,199 SNPs, the 28,544 remaining SNPs were used to construct the phylogenetic tree using PhyML v.3.0[16] using the HKY85 model with four categories for the gamma model of rate distributions and 100 bootstraps.

To examine possible associations between *M. tuberculosis* lineage and drug susceptibility, the lineage was determined for each of the 322 strains using a 62-SNP barcode[17]. The resulting classification in the main *M. tuberculosis* lineages also served as a quality check for the generated maximum-likelihood phylogenetic tree, as it enabled us to validate that isolates belonging to the same lineage clustered together in the tree.

Statistical analysis

The $\chi 2$ test was used to statistically test the association between *M. tuberculosis* lineage and drug susceptibility. Cohen's k was used to determine the level of agreement between WGS and phenotypic DST for first-line drugs. In addition, enrichment values were calculated for drug resistance per lineage based on the ratio of lineage-specific observed and expected occurrence of drug resistance. The ratios were visualized in a heat map as a measure of association between *M. tuberculosis* lineage and drug susceptibility.

Results

Patient characteristics

Patients were mostly young (median age 33 years, interquartile range 23.5–45.0 years), 51.9% were male and 16.1% reported a history of TB treatment. Using WGS, mutations associated with drug resistance to any TB drug were identified in 51 (15.8%) of the 322 patients; 42 (13.0%) were identified to have primary resistance, MDR-TB was present in 8 patients (2.5%), including primary MDR-TB in 5 patients (1.6%) and primary extensively drug-resistant TB (XDR-TB) in 1 patient (0.3%) (Table 2.1, Table S2.2).

Mutations associated with drug resistance

Genetic variants in multiple genes associated with drug resistance in *M. tuberculosis* were identified by WGS (Table 2.2). A total of 29 isolates (9.0%) had mutations in genes associated with resistance to INH, including *katG* (n = 18), *kasA* (n = 3) and the promoter region of *fabG1* C-15T (n = 9). Nine isolates with mutations at the *fabG1* promoter site were predicted to have co-resistance to INH and ethionamide. The most common mutation in *katG* was Ser315Thr (n = 14), but the uncommon mutation *katG* Ser315Met was also found in two isolates, and *katG* Trp191Arg and *katG* Ala106Val were found in one isolate each. The *katG* Ser315Thr mutation was found in seven of eight MDR-TB isolates (Table 2.2). RIF resistance-conferring mutations were identified in ten isolates (3.1%), all in the *rpoB* gene. Six common mutations distributed over codons 435, 445 and 450 were found, mostly as single mutations at one of these positions. Rare mutations at codon 432 and 441 were found together in one isolate. One MDR-TB isolate with *rpoB* Ser450Leu also harboured the well-known compensatory mutation at *rpoC* Leu527Val.

For EMB, nine isolates (2.8%) showed mutations in the *embB* gene, with the mutation Met 306Val being most common (n = 4), and rare mutations at codon 406 and codon 497 in three MDR isolates. PZA resistance-conferring mutations were identified in 18 isolates (5.6%). All mutations occurred as a single mutation in five codons in the *pncA*

gene. For fluoroquinolones (FQs), mutations in *gyrA* codon 90 and 94 were identified in 4 isolates (1.2%), three with FQ monoresistance and one XDR-TB isolate. With regard to resistance to STR and injectable agents, mutations were found in the *rpsL* and *rrs* loci in nine isolates (2.8%). Mutations at *rrs* codon 492 were only found in STR-monoresistant strains. The XDR-TB isolate harboured a mutation at the *eis* promoter linked to kanamycin resistance.

Table 2.1. Presence of drug resistance mutations in 322 clinical *M. tuberculosis* isolates detected by whole genome sequencing

Resistance pattern	Number (%) of strains
Susceptible to all drugs	271 (84.3)
Resistant to any drug	51 (15.8)
Resistant to first-line drugs	
Any first-line drug	48 (14.9)
Isoniazid	29 (9.0)
Rifampicin	10 (3.1)
Ethambutol	8 (2.5)
Pyrazinamide	18 (5.6)
Streptomycin	9 (2.8)
Resistant to second-line drugs	
Ethionamide	9 (2.8)
Fluoroquinolones	4 (1.2)
Amikacin	1 (0.3)
Kanamycin	1 (0.3)
Monoresistance	
Isoniazid	10 (3.1)
Rifampicin	1 (0.3)
Pyrazinamide	15 (4.7)
Streptomycin	3 (0.9)
Fluoroquinolones	3 (0.9)
Resistance to multiple drugs	
Multidrug resistance (MDR)	8 (2.5)
Extensively drug resistance (XDR)	1 (0.3)
Polyresistance*	10 (3.1)

*Defined as resistance to multiple drugs but not M/XDR

Table 2.2. Distribution of drug resistance-associated mutations in 51 *M. tuberculosis* isolates with any drug resistance identified by whole genome sequencing

Drug	Gene	Amino acid change	No. of isolates	No. of MDR-TB isolates
INH	*katG*	Ser315Thr	14	7[a]
		Ser315Met	2	0
		Trp191Arg	1	0
	fabG1	C-15T promoter	8	1
	kasA	Gly312Ser	3	0
	Double loci *fabG1* + *katG*	C-15T promoter + Ala106Val	1	1
RIF	rpoB	His445Tyr	1	1
		His445Asp	1	1
		Asp435Val	1	1
		Asp435Tyr	2	2
		Ser450Leu	2	2[a]
		His445Cys	1	1
		Gln432Lys + Ser441Leu	1	0
	Double loci *rpoB* + *rpoC*	Ser450Leu + Leu527Val	1	1
EMB	*embB*	Met306Ile	2	1[a]
		Met306Val	4	1
		Gly406Asp	1	1
		Gln497Lys	1	1
		Met306Val + Gly406Asp	1	1
STR	*rrs*	C492T	3	0
		A514C	1	1
	rpsL	Lys43Arg	5	3[a]
PZA	*pncA*	His82Arg	6	0
		Thr87Met	9	1
		Ser66Pro	1	1
		Pro62Leu	1	0
		Ala171Val	1	1[a]
ETO	*fabG1*	C-15T promoter	9	2
FLQ	*gyrA*	Asp94Gly	1	0
		Ala90Val	1	0
		Asp94Asn	1	0
		Asp94Ala	1	1[a]
AMK	*rrs*	A514C	1	1
KAN	*eis*	G-14A promoter	1	1[a]

*Abbreviations: INH: isoniazid; RIF: rifampicin; EMB: ethambutol; STR: streptomycin; PZA: pyrazinamide; ETO: ethionamide; FLQ: fluoroquinolones; AMK: amikacin; KAN: kanamycin; MDR-TB: multidrug-resistant.
[a]Mutation occurred in the extensively drug-resistant (XDR) isolate.

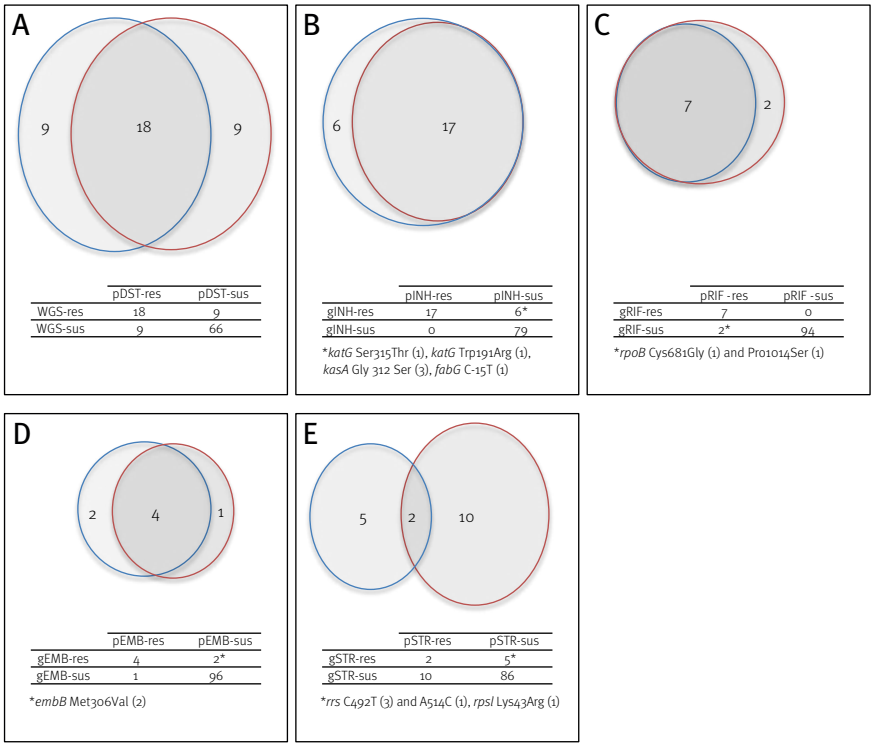
Agreement of phenotypic and genotypic drug susceptibility testing to first-line drugs

Overall, there was considerable agreement between genotypic and phenotypic DST. Phenotypic DST for the first-line drugs RIF, EMB and STR was available for 103 isolates with WGS data and for INH it was available for 102 isolates, with resistance found to INH (n = 17), RIF (n = 7), EMB (n = 6) and STR (n = 7). Concordance between WGS and phenotypic DST was high for RIF (k = 0.865; P < 0.001) and INH (k = 0.814; P < 0.001) but was low for EMB (k = 0.712; P < 0.001) and STR (k = 0.136; P = 0.148). Agreement of genotypic and phenotypic DST for each drug is shown in Figure 2.1.

The common mutations *katG* Ser315Thr and *fabG1* C-15T were highly predictive of phenotypic INH resistance in this setting (89.5% of the isolates with either one of these mutations were phenotypically resistant). Similarly, mutations in the RIF resistance-determining region (RRDR) of *rpoB* were highly predictive of resistance to RIF (100% of the isolates with a mutation were phenotypically resistant). Nine isolates showed drug resistance-associated mutations but were susceptible by phenotypic DST. In those isolates, mutations were found for STR at *rrs* C492T (n = 3) and A514C (n = 1) and *rpsL* Lys43Arg (n = 1); for INH at *katG* Ser315Thr (n = 1), *katG* Trp191Arg (n = 1), *fabG1* C-15T (n = 1) and *kasA* Gly312Ser (n = 3); and for EMB at *embB* Met306Val (n = 1) and Met306Ile (n = 1). Conversely, nine other isolates that were drug-resistant according to phenotypic DST (n = 7 for STR, n = 2 for RIF and n = 1 for EMB) showed no known drug resistance mutations using WGS. However, one isolate with phenotypic RIF resistance but without well-known drug resistance mutations harboured other mutations in *rpoB* (*rpoB* Cys681Gly and *rpoB* Pro1014Ser).

Phylogenetic distribution of drug resistance

Drug resistance rates and patterns differed significantly between different *M. tuberculosis* lineages (Figure 2.2; Table S2.3). Among 116 isolates belonging to the East-Asian lineage, 12 (10.3%) were genotypically drug-resistant compared with 36 (18.2%) of 198 Euro-American strains and 3 (37.5%) of 8 Indo-Oceanic strains ($\chi 2 = 6.258$; P = 0.044). Although fewer strains belonging to the East-Asian lineage had drug resistance mutations, they more often had multiple drug resistance mutations ($\chi 2 = 5.844$; P = 0.054). From the phylogeny, it was observed that most of the *pncA* mutations in isolates harbouring PZA resistance clustered together. The nine isolates with a *pncA* Thr87Met mutation were adjacent in the tree, and the same goes for the six isolates carrying the *pncA* His82Arg mutation, suggesting transmitted resistance. However, the isolates differed by more than 12 SNPs, the commonly used threshold for (recent) transmission [18]. The other three mutations occurred only once and in genetically distant strains.



	pDST-res	pDST-sus
WGS-res	18	9
WGS-sus	9	66

	pINH-res	pINH-sus
gINH-res	17	6*
gINH-sus	0	79

**katG* Ser315Thr (1), *katG* Trp191Arg (1), *kasA* Gly 312 Ser (3), *fabG* C-15T (1)

	pRIF -res	pRIF -sus
gRIF-res	7	0
gRIF-sus	2*	94

**rpoB* Cys681Gly (1) and Pro1014Ser (1)

	pEMB-res	pEMB-sus
gEMB-res	4	2*
gEMB-sus	1	96

**embB* Met306Val (2)

	pSTR-res	pSTR-sus
gSTR-res	2	5*
gSTR-sus	10	86

**rrs* C492T (3) and A514C (1), *rpsl* Lys43Arg (1)

Figure 2.1. Comparison of phenotypic drug susceptibility testing (DST) and DST by whole-genome sequencing (WGS) for *Mycobacterium tuberculosis* isolates to (A) any first-line drug, (B) isoniazid (INH), (C) rifampicin (RIF), (D) ethambutol (EMB) and (E) streptomycin (STR). Light grey, drug resistance determined by WGS; dark grey, drug resistance determined by phenotypic DST. Data are presented as number of strains. N = 103 for RIF, EMB and STR and N = 102 for INH.

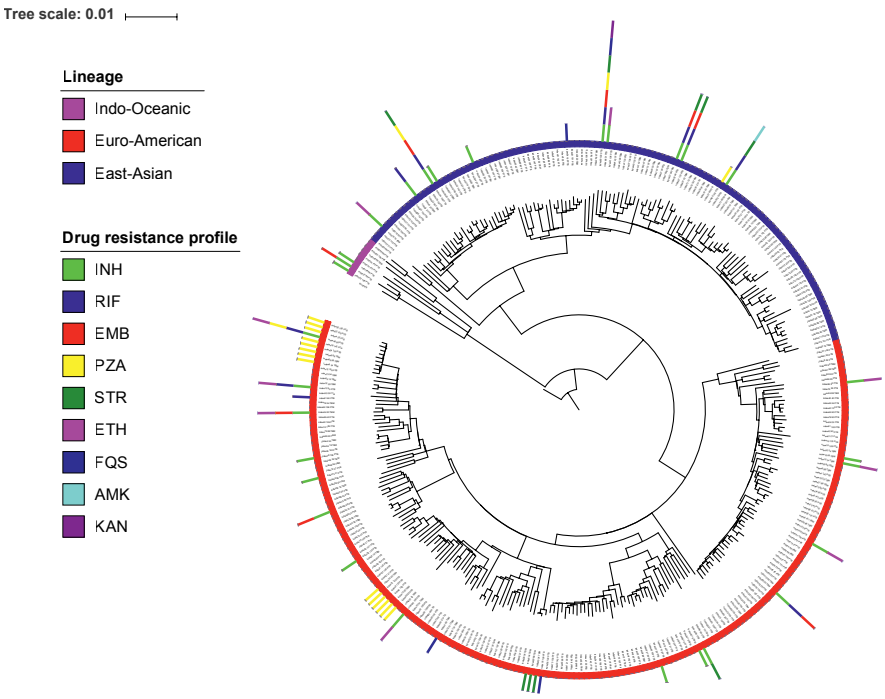


Figure 2.2. Phylogenetic tree of 322 *Mycobacterium tuberculosis* isolates indicating drug resistance profiles and main lineages based on single nucleotide polymorphism (SNP) barcoding. INH, isoniazid; RIF rifampicin; EMB, ethambutol; PZA, pyrazinamide; STR, streptomycin; ETH, ethionamide; FQS, fluoroquinolones; AMK, amikacin; KAN, kanamycin.

Discussion

This is the first study to report the use of WGS on clinical *M. tuberculosis* isolates from Indonesia, showing its potential for clinical management and TB control in Indonesia. Indonesia has a huge gap between MDR-TB incidence and MDR-TB treatment. Together with China, India, Nigeria and the Russian Federation, it accounted for >60% of this gap on a global scale in 2015[2]. Of the 10,000 estimated MDR- or RIF-resistant notified pulmonary TB cases in Indonesia in 2015, only an estimated 15% were started on treatment[2]. A large part of this gap can be explained by poor detection of drug resistance. Phenotypic DST is poorly accessible in remote parts of the country. WGS could offer a rapid and comprehensive diagnostic solution, especially with the introduction of portable platforms and the decreasing price and turnaround time of WGS[18], leading to quicker and more appropriate treatment.

In a random collection of patient isolates, it was found that 15.8% carried mutations associated with drug resistance to any TB drug, often in patients without a history of TB treatment. This is much lower than reported in an earlier study in Indonesia where resistance to first-line drugs was reported in 38.2% of 262 culture-positive samples, although these samples were collected over the whole country and resistance was determined by phenotypic DST[19]. In the current study, resistance mutations to all first-line drugs, FQs and second-line injectable drugs were found. MDR-TB was present in eight isolates (2.5%), which is lower than expected based on recent survey data[2], and XDR-TB was present in one isolate (0.3%).

In contrast to the findings from previous studies[19–21], in the current study drug resistance was observed to occur more often in strains belonging to the Indo-Oceanic lineage compared with the Euro-American and East-Asian lineages. On the other hand, strains belonging to the East-Asian lineage more often harboured multiple mutations. One possible explanation is that the East-Asian lineage, which includes Beijing strains, appears to have a higher mutation rate that could lead to accelerated acquisition of drug resistance mutations. It may be one of the reasons why this lineage has been repeatedly associated with drug resistance[22,23]. However, this would also lead to more resistance against a single drug.

Published studies from Indonesia have clearly shown disparities in the relative distribution of *M. tuberculosis* genotypes across the archipelago. In this article, we reported resistance mutations in a urban setting in Java where modern lineage 2 (East-Asian) and lineage 4 (Euro-American) are highly prevalent. These findings should be confirmed in other regions, especially eastern Indonesia, where lineage 1 (Indo-Oceanic) occurs at a substantially higher frequency[24–26]. Current catalogues of drug resistance mutations rely predominantly on data from modern lineages 2 and 4[10,27]. A recent study from India, where lineage 1 and 3 predominate, reported putative novel resistance-conferring mutations in lineage 1 that had not previously been implicated in resistance[27], suggesting that the mutations underlying genotypic drug resistance differ by lineage.

WGS offers the possibility to learn more about the influence of the genetic background of strains on all aspects of drug resistance evolution in *M. tuberculosis*. WGS could provide insight into secondary compensatory mutations, not conferring resistance but reducing the fitness cost of the resistance mutation by interacting epistatically with it. The rate of acquiring new drug resistance mutations and the fitness costs of these mutations may vary as a function of the strain genetic background[27,28].

Most mutations associated with INH resistance were found in *katG* (62.1%) and the promoter region of fabG1 (31.0%). Testing only for *katG* and *fabG1* promoter mutations would detect INH resistance in 89.6% of the INH-resistant isolates. All mutations associated with RIF resistance were found only in the RRDR of *rpoB*. This therefore confirms the usefulness of these three specific regions for prediction of INH and RIF resistance[29]. Conversely, *kasA* Gly312Ser and *katG* Trp191Arg mutations were found in phenotypically INH-susceptible isolates. *KasA* (β-ketoacyl ACP synthase), coded by the *kasA* gene, is an enzyme involved in mycolic acid synthesis. Mutations in the *kasA* gene have been associated with low-level INH resistance[30,31] but the role of these mutations in INH resistance remains unclear. Therefore, the significance of this gene in conferring drug resistance in Indonesia should be carefully assessed. Two isolates were phenotypically RIF-resistant but harboured mutations in *rpoB* that have not been associated with drug resistance before (Cys681Gly and Pro1014Ser), therefore their possible role in RIF resistance should be confirmed in other studies.

Regarding EMB, three loci of the *embB* gene (loci 306, 406 and 497) were identified, with mutations at locus 306 being the most frequent. Mutations in *embB* were present only in combination with other drug resistance mutations, in line with several studies that have demonstrated a strong association between embB306 mutations and INH-, RIF- or multidrug-resistant TB[32]. This finding suggests that embB306 mutations may have a selective advantage upon treatment with multiple drugs[33]. Regarding STR, isolates carried mutations in *rpsL* (n = 5) and *rrs* (n = 4). A common variant in *rpsL* (Lys43Arg) has been associated with high-level STR resistance[32] and this variant was indeed the most common in this study. Phenotypic STR resistance was only confirmed in two of seven genotypically STR-resistant isolates with DST results available. Three isolates had a single mutation in *rrs* codon 492; their relevance should be carefully interpreted since this mutation has been reported as a marker for the LAM3 genetic lineage of *M. tuberculosis* rather than for STR resistance[34].

In line with previous studies, concordance between WGS and phenotypic DST was good for INH and RIF[29] but had low agreement for EMB and STR[35,36]. There are several possible explanations for this finding. First, discordance was mainly found with uncommon genotypic mutations, which may be associated with low-level resistance that can be missed by conventional phenotypic DST. Second, the difference between the epidemiological breakpoint and the minimum inhibitory concentration (MIC) for EMB and STR is relatively small[35,36], which complicates phenotypic DST. Third, detection of phenotypic STR resistance in the absence of known mutations for STR resistance suggests the existence of other resistance mechanisms such as efflux pumps[36]. Finally, we cannot exclude errors in phenotypic DST despite good quality control in the laboratory. In this regard, the fact that one isolate was phenotypically

drug-susceptible with a high confidence katG Ser315Thr mutation for INH in WGS is a matter of concern since this mutation is associated with high-level resistance[29,37]. Given this high specificity, mutations at katG315 might even be used to assess laboratory quality for DST.

Second-line drug resistance was observed in a number of isolates. At present, there is no standard phenotypic DST for PZA and the second-line drugs in Indonesia, and evidence is limited on the performance of DST for these drugs[38], indicating once more the potential added value of WGS in this setting. Also, reported prior treatment for TB poorly corresponded with genotypic drug resistance, and primary drug resistance was common in this study population of Indonesian patients. Of concern, the only XDR-TB strain was found in a patient who had reported no prior TB treatment. This suggests that selective use of genotypic or phenotypic DST, targeting only those with previous treatment or other factors presumed to be associated with drug resistance, will result in many undetected resistant cases and ongoing transmission of drug-resistant strains[39]. This is further supported by the observation that 10 of 51 isolates with mutations associated with resistance to any drug were not phenotypically tested for drug susceptibility because they did not have these risk factors. WGS could help in early identification of resistance in these patients, limiting the spread of drug-resistant TB.

Molecular drug resistance for PZA mostly involves mutations in the *pncA* gene. There is no predominant drug resistance mutation in *pncA* but a range of diverse mutations across the gene, each associated with a different MIC[5]. Five different mutations in *pncA* were found in the isolates in this study, almost all in PZA-monoresistant strains. Most previous studies reported that PZA resistance is common in MDR-TB strains[40] although significant rates of PZA monoresistance have been reported in some settings[40–42]. A study in China showed that PZA monoresistance contributes to delay in the resolution of lung cavitation without affecting the sputum conversion and lesion elimination rates[41]. The non-essential nature of the *pncA* gene, such that it can accumulate various mutations without affecting the viability of the organism, might explain the spread of PZA-monoresistant strains. Two *pncA* mutations exclusively occurred in strains adjacent in the phylogenetic tree, suggesting possible transmission. However, the genetic difference between these strains was too large to conclude that this was indeed the case.

This study has several limitations. First, WGS was performed on a random sample of archived *M. tuberculosis* isolates so we cannot conclude that the proportions of drug resistance shown in this study are representative. This was a convenience sample collected from patients with pulmonary and meningeal TB. We estimate that several thousands of patients are treated for TB each year in Bandung. However, culture is not

routinely performed and isolates are not archived. Therefore, our sampling fraction is likely to be ‹10%. Second, phenotypic DST was available for most genotypically resistant strains but only for a fraction of isolates without resistance mutations. As a consequence, specificity estimation for genotypic resistance was not possible. Third, sequencing was performed retrospectively on archived isolates; therefore it was not possible to evaluate time to diagnosis of drug resistance using WGS. Nevertheless, we for the first time highlight the potential benefit of using WGS to generate an *in silico* drug susceptibility profile in Indonesia and show that mutations associated with drug resistance are highly predictive for phenotypic resistance to RIF and INH in the region. Larger studies are needed to confirm the clinical relevance of several uncommon mutations found in this strain collection. Given the fact that phenotypic DST is complex and slow and is poorly accessible in large parts of Indonesia, WGS could offer a rapid and comprehensive diagnostic solution. This technology is now more accessible with the introduction of portable platforms and an automatic bioinformatic pipeline as well as the decreasing price and turnaround time of WGS[43]. It could lead to quicker and more appropriate treatment in low-income settings where many still rely on empirical treatment regimens. This pilot study offers a good starting point to further evaluate the impact of WGS in the diagnosis, treatment, surveillance and control of drug-resistant TB in Indonesia.

Funding

This study is a result of collaborative projects between Indonesia and the Netherlands. Inclusion of patients and collection of *M. tuberculosis* isolates in the TB meningitis cohort were financially supported by the Royal Academy of Arts and Sciences (KNAW), Foundation for Advancement of Tropical Research (NWO-WOTRO), and IMPACT, a 5-year HIV programme supported by the European Commission. Inclusion of patients and collection of *M. tuberculosis* isolates in the pulmonary TB cohort were supported by TANDEM (TB and Diabetes Mellitus) from the European Community's Seventh Framework Programme [FP7/2007-2013] and Universitas Padjadjaran, Indonesia. Whole genome sequencing was financially supported by a VIDI grant from the Netherlands Organisation for Scientific Research [NWO 017.106.310], The Netherlands. LC has a fellowship from the Indonesian Ministry of Research, Technology and Higher Education (Indonesia) and Radboud university medical center (The Netherlands), and CR has a fellowship from Radboud university medical center. BED was supported by the Netherlands Organisation for Scientific Research [NWO VIDI grant 864.14.004]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests

None declared.

Ethical approval

Health Research Ethics Committee, Faculty of Medicine, Universitas Padjadjaran (Bandung, Indonesia).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jgar.2018.08.018.

References

1. Falzon D, Jaramillo E, Wares F, Zignol M, Floyd K, Raviglione MC. Universal access to care for multidrug-resistant tuberculosis: an analysis of surveillance data. Lancet Infect Dis 2013;13:690–7.
2. World Health Organization. Global tuberculosis report. 2016 http://apps.who. int/medicinedocs/en/d/Js23098en/. [Accessed 27 December 2018].
3. Lee RS, Behr MA. The implications of whole-genome sequencing in the control of tuberculosis. Ther Adv Infect Dis 2016;3:47–62.
4. Salamon H, Yamaguchi KD, Cirillo DM, Miotto P, Schito M, Posey J, et al. Integration of published information into a resistance-associated mutation database for *Mycobacterium tuberculosis*. J Infect Dis 2015;211(Suppl 2):S50–7.
5. Ramirez-Busby SM, Valafar F. Systematic review of mutations in pyrazinamidase associated with pyrazinamide resistance in *Mycobacterium tuberculosis* clinical isolates. Antimicrob Agents Chemother 2015;59:5267–77.
6. World Health Organization (WHO). Xpert MTB/RIF assay for the diagnosis of pulmonary and extrapulmonary TB in adults and children. Policy update. Geneva, Switzerland: WHO; 2013.
7. Steingart KR, Schiller I, Horne DJ, Pai M, Boehme CC, Dendukuri N. Xpert1 MTB/RIF assay for pulmonary tuberculosis and rifampicin resistance in adults. Cochrane Database Syst Rev 2014(1)CD009593.
8. Zetola NM, Shin SS, Tumedi KA, Moeti K, Ncube R, Nicol M, et al. Mixed *Mycobacterium tuberculosis* complex infections and false-negative results for rifampin resistance by GeneXpert MTB/RIF are associated with poor clinical outcomes. J Clin Microbiol 2014;52:2422–9.
9. Witney AA, Cosgrove CA, Arnold A, Hinds J, Stoker NG, Butcher PD. Clinical use of whole genome sequencing for *Mycobacterium tuberculosis*. BMC Med 2016;14:46.
10. Coll F, McNerney R, Preston MD, Guerra-Assuncao JA, Warry A, Hill-Cawthorne G, et al. Rapid determination of anti-tuberculosis drug resistance from whole genome sequences. Genome Med 2015;7:51.
11. van Laarhoven A, Dian S, Ruesen C, Hayati E, Damen MSMA, Annisa J, et al. Clinical parameters, routine inflammatory markers and LTA4H genotype as predictors for mortality among 608 tuberculous meningitis patients in Indonesia. J Infect Dis 2017;215:1029–39.
12. Chaidir L, Annisa J, Dian S, Moore DAJ, Muhsinin S, Parwati I, et al. MODS culture for primary diagnosis of tuberculous meningitis and HIV-associated pulmonary tuberculosis in Indonesia. Int J Trop Dis Health 2013;3:346–54.
13. Deatherage DE, Barrick JE. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. Methods Mol Biol 2014;1151:165–88.
14. Farhat MR, Shapiro BJ, Sheppard SK, Colijn C, Murray M. A phylogeny-based sampling strategy and power calculator informs genome-wide associations study design for microbial pathogens. Genome Med 2014;6:101.
15. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 2010;59:307–21.
16. Coll F, Preston M, Guerra-Assuncao JA, Hill-Cawthorn G, Harris D, Perdigao J, et al. PolyTB: a genomic variation map for *Mycobacterium tuberculosis*. Tuberculosis (Edinb) 2014;94:346–54.
17. Walker TM, Ip CLC, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. Lancet Infect Dis 2013;13:137–46.
18. Witney AA, Gould KA, Arnold A, Coleman D, Delgado R, Dhillon J, et al. Clinical application of whole-genome sequencing to inform treatment for multidrug resistant tuberculosis cases. J Clin Microbiol 2015;53:1473–83.
19. Lisdawati V, Puspandari N, Rif'ati L, Soekarno T, M Melatiwati, K Syamsidar, et al. Molecular epidemiology study of *Mycobacterium tuberculosis* and its susceptibility to anti-tuberculosis drugs in Indonesia. BMC Infect Dis 2015;15:366.
20. Black PA, de Vos M, Louw GE, van der Merwe RG, Dippenaar A, Streicher EM, et al. Whole genome sequencing reveals genomic heterogeneity and antibiotic purification in *Mycobacterium tuberculosis* isolates. BMC Genomics 2015;16:857.

21. Aung HL, Tun T, Moradigaravand D, Koser CU, Nyunt WW, Aung ST, et al. Whole-genome sequencing of multidrug-resistant *Mycobacterium tuberculosis* isolates from Myanmar. J Glob Antimicrob Resist 2016;6:113–7.
22. Fenner L, Egger M, Bodmer T, Altpeter E, Zwahlen M, Jaton K, et al. Effect of mutation and genetic background on drug resistance in *Mycobacterium tuberculosis*. Antimicrob Agents Chemother 2012;56:3047–53.
23. Trauner A, Borrell S, Reither K, Gagneux S. Evolution of drug resistance in tuberculosis: recent progress and implications for diagnosis and therapy. Drugs 2014;74:1063–72.
24. Chaidir L, Sengstake S, de Beer J, Krismawati H, Lestari FD, Ayawaila S, et al. *Mycobacterium tuberculosis* genotypic drug resistance patterns and clustering in Jayapura, Papua, Indonesia. Int J Tuberc Lung Dis 2015;19:428–33.
25. Chaidir L, Sengstake S, de Beer J, Oktavian A, Krismawati H, Muhapril E, et al. Predominance of modern *Mycobacterium tuberculosis* strains and active transmission of Beijing sublineage in Jayapura, Indonesia Papua. Infect Genet Evol 2016;39:187–93.
26. Sasmono RT, Massi MN, Setianingsih TY, Wahyuni S, Anita, Halik H, et al. Heterogeneity of *Mycobacterium tuberculosis* strains in Makassar, Indonesia. Int J Tuberc Lung Dis 2012;16:1441–8.
27. Gygli SM, Borrell S, Trauner A, Gagneux S. Antimicrobial resistance in *Mycobacterium tuberculosis*: mechanistic and evolutionary perspectives. FEMS Microbiol Rev 2017;41:354–73.
28. Manson AL, Abeel T, Galagan JE, Sundaramurthi JC, Salazar A, Gehrmann T, et al. *Mycobacterium tuberculosis* whole genome sequences from Southern India suggest novel resistance mechanisms and the need for region-specific diagnostics. Clin Infect Dis 2017;64:1494–501.
29. Rodwell TC, Valafar F, Douglas J, Qian L, Garfein RS, Chawla A, et al. Predicting extensively drug-resistant *Mycobacterium tuberculosis* phenotypes with genetic mutations. J Clin Microbiol 2014;52:781–9.
30. Unissa AN, Subbian S, Hanna LE, Selvakumar N. Overview on mechanisms of isoniazid action and resistance in *Mycobacterium tuberculosis*. Infect Genet Evol 2016;45:474–92.
31. Lee ASG, Lim IHK, Tang LLH, Telenti A, Wong SY. Contribution of kasA analysis to detection of isoniazid-resistant *Mycobacterium tuberculosis* in Singapore. Antimicrob Agents Chemother 1999;43:2087–9.
32. Jagielski T, Ignatowska H, Bakula Z, Dziewit L, Napiorkowska A, Augustynowicz-Kopec E, et al. Screening for streptomycin resistance-conferring mutations in *Mycobacterium tuberculosis* clinical isolates from Poland. PLoS One 2014;9:e100078.
33. Bakula Z, Napiorkowska A, Bielecki J, Augustynowicz-Kopec E, Zwolska Z, Jagielski T. Mutations in the embB gene and their association with ethambutol resistance in multidrug-resistant *Mycobacterium tuberculosis* clinical isolates from Poland. Biomed Res Int 2013;2013:167954.
34. Villellas C, Aristimuno L, Vitoria MA, Prat C, Blanco S, Garcia de Viedma D, et al. Analysis of mutations in streptomycin-resistant strains reveals a simple and reliable genetic marker for identification of the *Mycobacterium tuberculosis* Beijing genotype. J Clin Microbiol 2013;51:2124–30.
35. Horne DJ, Pinto LM, Arentz M, Lin SY, Desmond E, Flores LL, et al. Diagnostic accuracy and reproducibility of WHO-endorsed phenotypic drug susceptibility testing methods for first-line and second-line antituberculosis drugs. J Clin Microbiol 2013;51:393–401.
36. Martin LJ, Roper MH, Grandjean L, Gilman RH, Coronel J, Caviedes L, et al. Rationing tests for drug-resistant tuberculosis—who are we prepared to miss? BMC Med 2016;14:30.
37. Cui Z, Li Y, Cheng S, Yang H, Lu J, Hu Z, et al. Mutations in the embC–embA intergenic region contribute to *Mycobacterium tuberculosis* resistance to ethambutol. Antimicrob Agents Chemother 2014;58:6837–43.
38. Ssengooba W, Meehan CJ, Lukoye D, Kasule GW, Musisi K, Joloba ML, et al. Whole genome sequencing to complement tuberculosis drug resistance surveys in Uganda. Infect Genet Evol 2016;40:8–16.
39. Mokrousov I, Narvskaya O, Otten T, Limeschenko E, Steklova L, Vyshnevskiy B. High prevalence of KatG Ser315Thr substitution among isoniazid-resistant *Mycobacterium tuberculosis* clinical isolates from northwestern Russia, 1996 to 2001. Antimicrob Agents Chemother 2002;46:1417–24.
40. Parsons LM, Salfinger M, Clobridge A, Dormandy J, Mirabello L, Polletta VL, et al. Phenotypic and molecular characterization of *Mycobacterium tuberculosis* isolates resistant to both isoniazid and ethambutol. Antimicrob Agents Chemother 2005;49:2218–25.
41. Kurbatova EV, Cavanaugh JS, Dalton T, Click ES, Cegielski JP. Epidemiology of pyrazinamide-resistant tuberculosis in the United States, 1999–2009. Clin Infect Dis 2013;57:1081–93.

42. Tan S, Rao Y, Guo J, Tan Y, Cai X, Kuang H, et al. The influence of pyrazinamide monoresistance on treatment outcomes in tuberculosis patients from southern China. J Tuberc Res 2016;4:9–17.
43. Cheng SJ, Thibert L, Sanchez T, Heifets L, Zhang Y. pncA mutations as a major mechanism of pyrazinamide resistance in *Mycobacterium tuberculosis*: spread of a monoresistant strain in Quebec, Canada. Antimicrob Agents Chemother 2000;44:528–32.

Supplementary tables

Table S2.1. PE/PPE genes and drug resistance genes excluded for the phylogeny construction

PE/PPE genes and genes in repetitive regions						Known drug resistance genes
Rv0031	Rv0922	Rv1575	Rv2107	Rv2741	Rv3381c	
Rv0096	Rv0977	Rv1576c	Rv2108	Rv2768c	Rv3386	
Rv0109	Rv0978c	Rv1577c	Rv2123	Rv2769c	Rv3387	accD6
Rv0124	Rv0980c	Rv1578c	Rv2126c	Rv2770c	Rv3388	ahpC
Rv0151c	Rv1034c	Rv1579c	Rv2162c	Rv2791c	Rv3425	efpA
Rv0152c	Rv1035c	Rv1580c	Rv2167c	Rv2810c	Rv3426	embA
Rv0159c	Rv1036c	Rv1581c	Rv2168c	Rv2812	Rv3427c	embB
Rv0160c	Rv1039c	Rv1582c	Rv2177c	Rv2814c	Rv3428c	embC
Rv0256c	Rv1040c	Rv1583c	Rv2278	Rv2815c	Rv3429	embR
Rv0278c	Rv1041c	Rv1584c	Rv2279	Rv2853	Rv3430c	ethA
Rv0279c	Rv1042c	Rv1585c	Rv2328	Rv2885c	Rv3474	fabD
Rv0280	Rv1047	Rv1586c	Rv2340c	Rv2892c	Rv3475	fadE24
Rv0285	Rv1054	Rv1646	Rv2352c	Rv2943	Rv3477	fbpC
Rv0286	Rv1067c	Rv1651c	Rv2353c	Rv2943A	Rv3478	furA
Rv0297	Rv1068c	Rv1705c	Rv2354	Rv2944	Rv3507	gid
Rv0304c	Rv1087	Rv1706c	Rv2355	Rv2961	Rv3508	gyrA
Rv0305c	Rv1088	Rv1753c	Rv2356c	Rv2978c	Rv3511	gyrB
Rv0335c	Rv1089	Rv1756c	Rv2371	Rv3018A	Rv3512	inhA
Rv0354c	Rv1091	Rv1757c	Rv2396	Rv3018c	Rv3514	iniA
Rv0355c	Rv1135c	Rv1763	Rv2408	Rv3021c	Rv3532	iniB
Rv0387c	Rv1149	Rv1764	Rv2424c	Rv3022A	Rv3533c	iniC
Rv0388c	Rv1168c	Rv1765A	Rv2430c	Rv3022c	Rv3539	kasA
Rv0442c	Rv1169c	Rv1768	Rv2431c	Rv3023c	Rv3558	katG
Rv0453	Rv1172c	Rv1787	Rv2479c	Rv3115	Rv3590c	fabG1
Rv0532	Rv1195	Rv1788	Rv2480c	Rv3125c	Rv3595c	manB
Rv0578c	Rv1196	Rv1789	Rv2487c	Rv3135	Rv3621c	ndh
Rv0741	Rv1199c	Rv1790	Rv2490c	Rv3136	Rv3622c	nat
Rv0742	Rv1214c	Rv1791	Rv2512c	Rv3144c	Rv3636	oxyR
Rv0746	Rv1243c	Rv1800	Rv2519	Rv3159c	Rv3637	pncA
Rv0747	Rv1313c	Rv1801	Rv2591	Rv3184	Rv3638	rmlD
Rv0754	Rv1325c	Rv1802	Rv2608	Rv3185	Rv3640c	rpoB

Table S2.1. Continued

PE/PPE genes and genes in repetitive regions						Known drug resistance genes
Rv0755A	Rv1361c	Rv1803c	Rv2615c	Rv3186	Rv3650	
Rv0755c	Rv1369c	Rv1806	Rv2634c	Rv3187	Rv3652	
Rv0795	Rv1370c	Rv1807	Rv2646	Rv3191c	Rv3653	rpsL
Rv0796	Rv1386	Rv1808	Rv2648	Rv3325	Rv3738c	rrs
Rv0797	Rv1387	Rv1809	Rv2649	Rv3326	Rv3739c	Rv0340
Rv0832	Rv1396c	Rv1818c	Rv2650c	Rv3327	Rv3746c	Rv1592c
Rv0833	Rv1430	Rv1840c	Rv2651c	Rv3343c	Rv3751	Rv1772
Rv0834c	Rv1441c	Rv1917c	Rv2652c	Rv3344c	Rv3798	Rv2242
Rv0850	Rv1450c	Rv1918c	Rv2653c	Rv3345c	Rv3812	Rv3124
Rv0872c	Rv1452c	Rv1983	Rv2654c	Rv3347c	Rv3827c	Rv3125c
Rv0878c	Rv1468c	Rv2013	Rv2655c	Rv3348	Rv3844	Rv3126c
Rv0915c	Rv1548c	Rv2014	Rv2656c	Rv3349c	Rv3872	thyA
Rv0916c	Rv1573	Rv2105	Rv2657c	Rv3350c	Rv3873	tlyA
Rv0920c	Rv1574	Rv2106	Rv2659c	Rv3367	Rv3892c	accD6
			Rv2666	Rv3380c	Rv3893c	

Table S2.2. Resistance profile for 51 strains with any drug resistance determined by WGS

Study number	Pheno-type	WGS INH	WGS RIF	WGS EMB	WGS STR	WGS MDR	Phen INH	Phen EMB	Phen STR	Phen MDR	Hist TB tr.	Lineage
100884	PTB	0	0	0	0	0	0	0	0	0	0	Euro-American (LAM)
100890	PTB	1	0	0	0	0	1	0	0	0	0	Euro-American (mainly T)
100913	PTB	1	0	0	0	0	1	0	0	0	0	East-Asian (Beijing)
100921	PTB	0	0	0	0	0	0	0	0	0	0	Euro-American (mainly T)
100926	TBM	1	1	0	0	1	1	0	0	1	0	East-Asian (Beijing)
100962	TBM	1	1	0	1	1	1	0	0	1	1	East-Asian (Beijing)
100964	TBM	1	0	0	0	0	1	0	1	0	0	Euro-American
100972	TBM	1	0	0	0	0	1	0	0	0	0	East-Asian (Beijing)
100989	TBM	1	0	1	0	0	1	0	0	0	0	East-Asian (Beijing)
1100217	TBM	1	1	1	1	1	1	1	1	0	0	East-Asian (Beijing)
1100232	TBM	1	0	1	0	0	0	0	0	0	0	Indo-Oceanic
1100238	TBM	1	0	1	0	0	1	1	0	0	0	Euro-American
1100294	TBM	0	1	0	0	0	0	0	1	0	0	Euro-American (mainly T)
1100607	TBM	1	0	0	0	0	0	0	0	0	0	Indo-Oceanic
1100707	TBM	0	0	0	0	0	0	0	0	0	0	Euro-American (LAM)
1100773	TBM	0	0	0	0	0	0	0	0	0	0	Euro-American (mainly T)
1100797	TBM	1	0	0	0	0	0	0	0	0	1	Indo-Oceanic
1100903	TBM	1	1	1	1	1				1	0	East-Asian (Beijing)
1100969	TBM	1	0	0	0	0					1	Euro-American (Cameroon)
1101011	TBM	1	0	0	0	0	0	0	0	0	0	Euro-American (mainly T)
1101025	PTB	1	0	0	1	0	1	0	0	0	0	Euro-American
1101028	PTB	1	1	0	0	1				1	0	Euro-American (mainly T)
1101054	PTB	0	0	0	0	0	0	0	1	0	0	Euro-American
1101303	PTB	1	1	0	0	1	1	0	0	1	1	Euro-American (mainly T)

1101323	TBM	1	0	0	0	0	1	0	0	0	1	Euro-American (X-type)
1101355	TBM	0	0	0	0	0	0	0	1	0	0	Euro-American (LAM)
1101572	TBM	0	0	0	0	0	0	0	0	0	0	East-Asian (Beijing)
1101628	TBM	0	0	0	0	0	0	0	0	0	0	Euro-American (mainly T)
900077	TBM	0	0	0	0	0	0	0	0	0	1	Euro-American (LAM)
900130	TBM	0	0	0	1	0	0	0	0	0	0	Euro-American (LAM)
900138	TBM	0	0	0	0	0	0	0	0	0	1	Euro-American (LAM)
900161	TBM	1	0	0	0	0	1	0	0	0	0	Euro-American (X-type)
900164	TBM	1	1	1	0	1	1	1	1	1	1	Euro-American (X-type)
900174	TBM	0	0	0	1	0	0	0	0	0	0	Euro-American (LAM)
900178	TBM	0	0	0	0	0	0	0	0	0	1	Euro-American (mainly T)
900181	TBM	0	0	0	1	0	0	0	0	0	0	Euro-American (LAM)
900202	TBM	0	0	0	0	0	0	0	0	0	0	Euro-American (LAM)
900221	TBM	0	0	0	0	0	0	0	1	0	0	Euro-American (mainly T)
900229	TBM	1	0	0	0	0	1	0	0	0	0	Euro-American (X-type)
900319	TBM	1	1	1	1	1	1	1	0	1	0	East-Asian (Beijing)
900365	PTB	0	0	0	0	0	0	0	0	0	0	Euro-American (mainly T)
900387	PTB	1	0	0	0	0	0	0	0	0	0	Euro-American
900603	PTB	1	0	1	0	0	1	0	0	0	0	Euro-American (mainly T)
900611	PTB	1	0	0	0	0	0	0	0	0	0	East-Asian (Beijing)
1100246	PTB	1	0	0	0	0					0	Euro-American (LAM)
1100353	PTB	1	0	0	0	0					0	Euro-American (LAM)
1100907	PTB	0	0	0	0	0					0	Euro-American (LAM)
1100943	PTB	0	0	0	0	0					0	East-Asian (Beijing)
1100995	PTB	0	0	0	0	0					0	Euro-American (mainly T)
1101933	PTB	0	0	0	0	0					0	Euro-American (mainly T)
1100909	PTB	1	1	1	1	1					0	East-Asian (Beijing)

'1' indicates genotypic (WGS) or phenotypic (Phen) drug resistance detected, 'o' no genotypic (WGS) or phenotypic (Phen) drug resistance detected, an empty cell indicates the data are missing. PTB: pulmonary TB; TBM: tuberculous meningitis; WGS: whole genome sequencing; Phen: phenotypic drug susceptibility.

Table S2.3. Heatmap of observed vs. expected genotypic drug resistance ratios per lineage and gene

	East-Asian	Euro-American	Indo-Oceanic
katG	1,23	0,90	0,00
fabG1	0,62	1,26	0,00
kasA	0,00	0,00	40,25
rpoB	1,51	0,74	0,00
rpoC	2,78	0,00	0,00
embB	1,54	0,54	4,47
pncA	0,46	1,36	0,00
rrs	0,69	1,22	0,00
rpsL	2,22	0,33	0,00
gyrA	1,39	0,81	0,00
eis	2,78	0,00	0,00

Colours range from green (‹1, observed drug resistance lower than expected drug resistance) to red (›1, observed drug resistance higher than expected drug resistance).

3

Linking minimum inhibitory concentrations to whole genome sequence-predicted drug resistance in *Mycobacterium tuberculosis* strains from Romania

Carolien Ruesen*, Anca Lelia Riza*, Adriana Florescu, Lidya Chaidir, Cornelia Editoiu, Nicole Aalders, Dragos Nicolosu, Victor Grecu, Mihai Ioana, Reinout van Crevel, Jakko van Ingen.

* Equal contribution

Scientific Reports. 2018; 8:9676.

Abstract

Objectives

Mycobacterium tuberculosis drug resistance poses a major threat to tuberculosis control. Current phenotypic tests for drug susceptibility are time-consuming, technically complex, and expensive. Whole genome sequencing is a promising alternative, though the impact of different drug resistance mutations on the minimum inhibitory concentration (MIC) remains to be investigated.

Methods

We examined the genomes of 72 phenotypically drug-resistant *Mycobacterium tuberculosis* isolates from 72 Romanian patients for drug resistance mutations. MICs for first- and second-line drugs were determined using the MycoTB microdilution method. These MICs were compared to macrodilution critical concentration testing by the Mycobacterium Growth Indicator Tube (MGIT) platform and correlated to drug resistance mutations.

Results

Sixty-three (87.5%) isolates harboured drug resistance mutations; 48 (66.7%) were genotypically multidrug-resistant. Different drug resistance mutations were associated with different MIC ranges; *katG* S315T for isoniazid, and *rpoB* S450L for rifampicin were associated with high MICs. However, several mutations such as in *rpoB*, *rrs* and *rpsL*, or *embB* were associated with MIC ranges including the critical concentration for rifampicin, aminoglycosides or ethambutol, respectively.

Conclusions

Different resistance mutations lead to distinct MICs, some of which may still be overcome by increased dosing. Whole genome sequencing can aid in the timely diagnosis of *Mycobacterium tuberculosis* drug resistance and guide clinical decision-making.

Introduction

Drug-resistance poses a major threat to tuberculosis (TB) control. Global surveillance data suggest that in 2015, there were an estimated 480.000 new cases of multi-drug-resistant TB (MDR-TB), i.e. resistant to at least isoniazid (INH) and rifampicin (RIF), and that 9.5% of these cases were extensively drug-resistant (XDR), i.e. also resistant to amikacin (AMI), kanamycin (KAN), or capreomycin (CAP), and at least one fluoroquinolone (FLQ)[1]. It is estimated that less than half of these cases are detected, and an even smaller proportion receive appropriate treatment[1]. One of the main problems in the control of drug-resistant TB (DR-TB) is the lack of laboratory capacity to diagnose resistance[2]. Conventional culture-based drug susceptibility testing (DST) is expensive, time-consuming and requires a specialized biosafety laboratory. Methods for DST for many of the second-line drugs have not yet been fully standardized[3]. In addition, these phenotypic methods often rely on growth of mycobacterial cultures in the presence of a single 'critical' drug concentration to distinguish resistant and susceptible strains based on epidemiological breakpoints, which are near the wild type minimum inhibitory concentration (MIC) distribution for some drugs[4].

Molecular assays are now available and can overcome some of the disadvantages of phenotypic methods to diagnose DR-TB. Yet, these assays are expensive and cannot identify all of the genetic loci associated with drug resistance[5,6]. Whole genome sequencing (WGS) of *M. tuberculosis*, tackles this problem and could enable the reliable prediction of the drug susceptibility phenotype within a clinically relevant timeframe. A fundamental issue in the application of WGS to predict drug resistance is how to interpret mutations in relation to phenotypic antibiotic resistance. The level of resistance, reflected by the MIC, a single nucleotide polymorphism (SNP) causes is most important for clinicians treating patients, in order to determine whether to increase the dosage or change the regimen.

Therefore, in the current study we provide this missing link by comparing MICs to drug resistance mutations determined by WGS. For this purpose we used a collection of drug-resistant strains from Romania, which had 20,000 new TB cases, and 800 new MDR-TB cases in 2012, making it one of the 18 high-priority countries for TB control in the World Health Organization (WHO) European Region[7].

Methods

Selection of *M. tuberculosis* isolates

We used *M. tuberculosis* isolates from the bio-archive established at the TB laboratory of the Clinical Hospital of Infectious Diseases and Pneumophtisiology "Victor Babes", Craiova, between 2014 and 2016. The laboratory runs approximately 7,000 sputum samples per year, covering much of the TB epidemic in Dolj county. Routine diagnostics as indicated by the National Control Programme include sputum microscopy and culture on Löwenstein-Jensen medium (liquid media introduced in May 2015). After primary culture on Löwenstein-Jensen medium, screening for INH and RIF-resistance is performed using INH 0.2 mg/l and RIF 40 mg/l-supplemented Löwenstein-Jensen slants.

All 73 isolates that were phenotypically resistant to INH and/or RIF were selected for the purpose of the present study. The 73 isolates were from different patients, and no serial isolates were included. Isolates stored in trypticase soy broth supplemented with glycerol were sub-cultured on Löwenstein-Jensen medium. Sub-cultured colonies were used for DNA extraction, and susceptibility testing.

Phenotypic drug susceptibility testing

All isolates were simultaneously subjected to MIC determinations of rifampicin, rifabutin, ethambutol, streptomycin, amikacin, kanamycin, ofloxacin, moxifloxacin, para-aminosalicylic acid (PAS), ethionamide and cycloserine by broth microdilution in Middlebrook 7H9 broth (SensiTitre MycoTB assay; Trek Diagnostics/ThermoFisher, Landsmeer, the Netherlands). Every isolate was tested once, unless contamination or no growth was observed; then the isolate was re-cultured and once more subjected to MIC testing. The MycoTB assay is a new method for susceptibility testing of *M. tuberculosis* complex[8,9]. The plate uses a 96-well microtiter broth format and contains 12 lyophilized first- and second-line antimycobacterial drugs. In contrast to other *M. tuberculosis* complex susceptibility methods, which test one or two critical concentrations of a drug, the MycoTB assay examines a range of drug concentrations and produces an MIC result. In addition, critical concentration testing for streptomycin, isoniazid, rifampicin and ethambutol (SIRE assay; BD BioScience, Erembodegem, Belgium) was performed using the Mycobacterium Growth Indicator Tube (MGIT) macrodilution platform (BD BioScience, Erembodegem, Belgium). The critical concentrations used for MGIT testing were 0.10 µg/ml for INH, 1.00 µg/ml for RIF, 5.00 µg/ml for EMB, and 1.00 µg/ml for STR. Technicians performing susceptibility tests were blinded to the sequencing results.

Whole genome sequencing and *in silico* determination of drug resistance

DNA isolation using UltraClean® Microbial DNA Isolation Kit (MO BIO Laboratories Inc., Carlsbad, CA) was followed by quantitative (Qubit v3.0) and qualitative assessment of the DNA (gel electrophoresis). We used the Nextera XT DNA Library Preparation Kit (Illumina, San Diego, CA) for library preparation. *M. tuberculosis* DNA was sequenced on an Illumina NextSeq500 instrument using 2 x 150 bp paired-end reads. One isolate was excluded due to failed library preparation, resulting in a total set of 72 isolates. No internal control was performed for WGS. After sequencing, the raw FASTQ sequence reads were filtered using Trim Galore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore), including removing of adapter sequences. Sequencing coverage was determined using the FASTQC quality control tool version 0.10.1, and the Genome Analysis Toolkit[10]. The average proportion of bases sequenced with a sequencing error rate of 0.1% or less per base was 77.8% per genome. The average coverage depth for the 72 sequenced strains was 165.1, and the average percentage of bases covered by at least one read was 99.5%. Quality control statistics are shown in Table S3.1.

Raw FASTQ sequencing files were uploaded to TB Profiler version 0.2.1, an online tool to determine drug resistance *in silico*[11]. It uses raw sequence data as input, aligns these to the *M. tuberculosis* H37Rv reference genome, and compares identified single nucleotide polymorphisms (SNPs) and indels to a curated list of 1,325 drug resistance mutations. Although it allows examining heteroresistance, we focused on majority variants. In addition, it determines the *M. tuberculosis* lineage based on a 62-SNP barcode[12]. The TB Profiler-predicted resistance mutations were validated using PhyResSE, version 1.0[13], another online tool that maps raw sequencing reads to the *M. tuberculosis* H37Rv reference strain to report drug resistance and phylogenetic SNPs. We used the *M. tuberculosis* complex numbering system, based on the sequence of the reference strain: H37Rv[14].

Phylogeny construction

A phylogeny was constructed to examine clustering of the isolates. The sequence reads were aligned to reference strain *M. tuberculosis* H37Rv, accession number NC_000962.3, and variants were called using Breseq software, version 0.27.1[15]. We extracted all 5,980 variable positions across the 72 *M. tuberculosis* sequences and concatenated them into a single alignment. Solely for the purpose of creating the phylogenetic tree, SNPs occurring in PE/PPE genes, genes related to mobile elements, as well as genes previously associated with drug resistance[11] were removed. The remaining 5,794 SNPs were used to construct the phylogenetic tree using PhyML, version 3.0[16] using the HKY85 model with four categories for the gamma distribution, and using a hundred bootstraps.

Comparison of MGIT, minimum inhibitory concentrations and whole genome sequencing

For each isolate, susceptibility to STR, INH, RIF, and EMB as determined by MGIT was compared to the MICs determined by MycoTB. If MycoTB MICs were higher than the critical concentration (›0.25 µg/ml for INH, ›1.0 µg/ml for RIF, ›4.0 µg/ml for EMB, and ›2.0 µg/ml for STR[17], the isolate was considered resistant to the respective drug. Cohen's kappa statistic was calculated for each drug to determine the level of agreement between MycoTB- and MGIT-determined resistance. In addition, drug resistance mutations identified by TB Profiler and/or PhyResSE were correlated to MGIT results and MICs. We visualized the MIC for each drug in histograms, and subsequently determined the frequencies of the individual resistance-determining mutations and plotted their distributions against the MGIT results for INH, RIF, EMB, and STR, and against the MICs for first- and second-line drugs. For isolates with discrepant susceptibility results for INH or RIF between WGS and MGIT, we compared the MGIT to the DST performed on Löwenstein-Jensen and in case of genotypically resistant, phenotypically susceptible isolates we examined the quality of the variant call (Table S3.2). All analyses were performed in R, version 3.3.1 (http://www.R-project.org).

Data availability

The raw sequence files (FASTQ) were archived on the NCBI Sequence Read Archive and are available at: http://www.ncbi.nlm.nih.gov/bioproject/475771. The individual isolates can be accessed under the following Biosample accession numbers: SAMN09402650-SAMN09402721. The Bioproject accession number is: PRJNA475771.

Results

Phenotypic drug susceptibility testing

MGIT results for INH, STR, and EMB were available for 68 out of 72 isolates (four cultures being contaminated); RIF MGIT results were available for 69 isolates, with three cultures contaminated. Fifty-seven isolates were phenotypically resistant to INH, 34 to RIF, 24 to STR, and three to EMB. Using the MycoTB microdilution method, MICs for INH, RIF, STR, EMB, ETO, OFL, MXF, PAS, cycloserine, AMK, and KAN were successfully determined for 64 isolates; six samples were contaminated, and two repeatedly failed to grow in the MycoTB plate. By MycoTB, 51 isolates were resistant to INH, 31 to RIF, nine to STR, four to EMB, 40 to ETO, four to OFL, one to MXF, one to PAS, one to cycloserine, two to AMK, and three were resistant to KAN (Figure S3.1). Poly-drug resistance was observed in 47, and MDR-TB in 28 isolates, no isolates were XDR-TB according to MycoTB. The agreement between MGIT and MycoTB for INH, RIF, STR and EMB was greater than to be expected by chance (Table S3.3). Cohen's kappa indicated

excellent agreement for INH ($\kappa = 0.794$; $p < 0.001$) and RIF ($\kappa = 0.905$; $p < 0.001$), and little agreement for STR ($\kappa = 0.289$; $p = 0.006$) and EMB ($\kappa = 0.246$; $p = 0.045$).

Whole genome sequencing and *in silico* determination of drug resistance

Sixty-three (87.5%) isolates harboured drug resistance-conferring mutations, and 48 (66.7%) had INH and RIF resistance-conferring mutations and were genotypically multidrug-resistant. The identified drug resistance mutations are listed in Table 3.1. Resistance-conferring mutations to INH were identified in 61 (84.7%) isolates.

The most common INH resistance mutation was *katG* S315T, which was present in 38 (52.8%) isolates as a single mutation, and in 16 (22.2%) together with a mutation in the *fabG1* promotor region (synonym: *inhA* promotor region). Resistance-conferring mutations to RIF were found in 50 (69.4%) isolates. The most prevalent RIF resistance-conferring mutation was *rpoB* S450L, observed in 17 (23.6%) isolates as a single mutation, and in one isolate as a combined mutation. Mutations in genes related to EMB susceptibility were identified in 39 (54.2%) isolates; *embB* M306I was observed in 25 (34.7%) isolates, and in 14 of these as a single mutation. Resistance to STR due to mutations in the *rpsL* and *rrs* loci was found in 19 (26.4%) isolates; with *rrs* A1401G in 13 isolates being the most frequent. This mutation however is not considered causative of STR resistance according to expert knowledge[18,19]. Mutations in the *pncA* gene, encoding the target of PZA were found in 23 (31.9%) isolates, 14 of which had a single *pncA* A146V mutation. Mutations to second-line drugs were less frequently observed, the *rrs* A1401G mutation associated with resistance to AMK, CAP, and KAN was most common, together with the C15T mutation in the promotor region of *fabG1* conferring resistance to ETO that was found in 20 (27.8%) isolates. The latter mutation is also associated with INH resistance. Only five isolates had one of four different mutations conferring resistance to FLQs.

Phylogenetic relatedness of the isolates

A phylogenetic tree was constructed based on 5,794 variable common nucleotide positions among the 72 *M. tuberculosis* isolates. Seventy-one isolates belonged to the Euro-American lineage and one belonged to the East-Asian lineage. The phylogeny showed that several isolates were closely related, but did not all harbour the same resistance-conferring mutations (Figure S3.2). Fourteen isolates within the same clade carried the *fabG1* (C-15T promotor), *katG* (S315T), *rpoB* (S450L), *embB* (M306I), and *pncA* (A146V) mutations, but they also harboured different mutations. Another 21 isolates clustered together in the tree and carried the S315T *katG* mutation, twenty of which had the H445N *rpoB* mutation, but apart from that their resistance profile differed. In addition, there were five pairs of neighbouring isolates that had similar resistance mutations.

Table 3.1. Frequency of *M. tuberculosis* resistance-conferring mutations

Drug	Gene	Mutation	Frequency (%)	MGIT-resistant[#]	MIC > CC[#]
INH	*katG*	**Ser315Thr**	**38 (52.8)**	**35/35**	**30/30**
	katG + *ahpC*	Asp311Gly + G-48A promoter	1 (1.4)	1/1	1/1
	katG + *fabG1*	Ser315Thr + C-15T promoter	16 (22.2)	15/15	16/16
	fabG1	C-15T promoter	4 (5.6)	4/4	2/4
	fabG1 + *inhA*	C-15T promoter + Ile21Val	1 (1.4)	1/1	0/1
	ahpC	C-52T promoter	1 (1.4)	1/1	1/1
RIF	*rpoB*	Asp435Val	4 (5.6)	3/4	3/4
		Gln432Pro	1 (1.4)	1/1	1/1
		His445Asn	12 (16.7)	2/11	1/7
		His445Asp	1 (1.4)	1/1	1/1
		His445Tyr	4 (5.6)	4/4	4/4
		Leu430Pro	2 (2.8)	0/1	0/1
		Ser450Leu	**17 (23.6)**	**15/15**	**16/16**
		Ser450Trp	1 (1.4)	1/1	1/1
		Asp435Gly + His445Asn	1 (1.4)	1/1	1/1
		Asp435Val + His445Asn	1 (1.4)	1/1	NA
		His445Asn + Pro454Leu	2 (2.8)	0/2	1/2
		Leu430Pro + His445Asn	1 (1.4)	1/1	NA
		Met434Ile + His445Asn	1 (1.4)	1/1	1/1
		Phe424Leu + His445Asn	1 (1.4)	1/1	0/1
		His445Asn + Ser450Leu + Pro-454Leu	1 (1.4)	1/1	0/1
EMB	*embA*	C-12T promoter	1 (1.4)	0/1	0/1
	embA + *embB*	C-12T promoter + Asp354Ala	1 (1.4)	0/1	0/1
		C-12T promoter + Glu504Asp	1 (1.4)	0/1	0/1
		C-12T promoter + Met306Ile	1 (1.4)	0/1	0/1
		C-12T promoter + Met306Leu + Asp354Ala	1 (1.4)	0/1	0/1
	embB	Asp354Ala	1 (1.4)	0/1	0/1
		Gly406Asp	5 (6.9)	0/5	0/1
		Met306Ile	**14 (19.4)**	**0/12**	**0/14**
		Met306Leu	1 (1.4)	0/1	0/1
		Met306Val	2 (2.8)	0/2	0/1
		Met306Ile + Gly406Asp	9 (12.5)	3/9	4/8
		Met306Val + Asp328Tyr	1 (1.4)	0/1	0/1
		Met306Ile + Tyr319Ser + As-p354Ala	1 (1.4)	0/1	0/1
STR	*rpsL*	Lys43Arg	1 (1.4)	1/1	1/1
		Lys88Arg	1 (1.4)	1/1	1/1
	rpsL + *rrs*	Lys43Arg + A1401G	2 (2.8)	2/2	2/2
	rrs	**A1401G***	**11 (15.3)**	**6/10**	**1/10**
		A514C	2 (2.8)	2/2	2/2
		C1402T*	1 (1.4)	1/1	0/1
		C517T	1 (1.4)	NA	0/1

Table 3.1. Continued

Drug	Gene	Mutation	Frequency (%)	MGIT-resistant#	MIC > CC#
PZA	*pncA*	**Ala146Val**	**14 (19.4)**		
		Gln10Arg	1 (1.4)		
		Gln10Stop	4 (5.6)	NA	NA
		Gly17Asp	2 (2.8)		
		Tyr34Stop	1 (1.4)		
		Tyr34Stop + Gly17Asp	1 (1.4)		
ETO	*ethA*	Thr61Met	1 (1.4)		1/1
	fabG1	**C-15T promoter**	**20 (27.8)**	NA	**19/19**
	fabG1 + *inhA*	C-15T promoter + Ile21Val	1 (1.4)		1/1
FQ	*gyrA*	Ala90Val	1 (1.4)		NA
		Asp94Gly	**2 (2.8)**		**1/2 (OFL);** 0/2 (MXF)
		Ser91Pro	1 (1.4)	NA	0/1 (OFL); 0/1 (MXF)
	gyrB	Asp461Asn	1 (1.4)		1/1 (OFL); 0/1 (MXF)
AMK	*rrs*	**A1401G**	**13 (18.1)**		**2/12**
		A514C*	2 (2.8)	NA	0/2
		C1402T	1 (1.4)		0/1
		C517T*	1 (1.4)		0/1
CAP	*rrs*	**A1401G**	**13 (18.1)**	NA	NA
		C1402T	1 (1.4)		
KAN	*eis*	G-14A promoter	2 (2.8)		1/2
	rrs	**A1401G**	**13 (18.1)**	NA	**1/12**
		C1402T	1 (1.4)		0/1

*These mutations possibly have no causative role in conferring drug resistance to the respective drug (18, 19). #Presented as N/N tested, because not all isolates with genotypic DST results had phenotypic DST results. The most frequently observed drug resistance mutations per drug are shown in bold. MGIT: Mycobacterium Growth Indicator Tube; MIC: minimum inhibitory concentration; CC: critical concentration; NA: not assessed.

Comparison of MGIT, minimum inhibitory concentrations and whole genome sequencing

First, whole genome sequencing was compared with phenotypic DST by MGIT. Figure S3.3 and Table 3.1 show the comparison of MGIT and WGS resistance prediction. For MGIT, all isolates with well-known INH resistance-conferring mutations were phenotypically resistant to INH, and all isolates without these mutations were susceptible to INH. For RIF, 13/69 (19%) discrepancies were observed between MGIT and WGS; all concerned WGS-resistant, MGIT-susceptible isolates. Eleven of these had a H445N *rpoB* mutation; eleven were susceptible according to DST performed on Löwenstein-Jensen medium, and variant call qualities were good (Table S3.2). For STR, WGS and MGIT did not match for 15/68 (22%) isolates; four had an A1401G *rrs* mutation but were susceptible

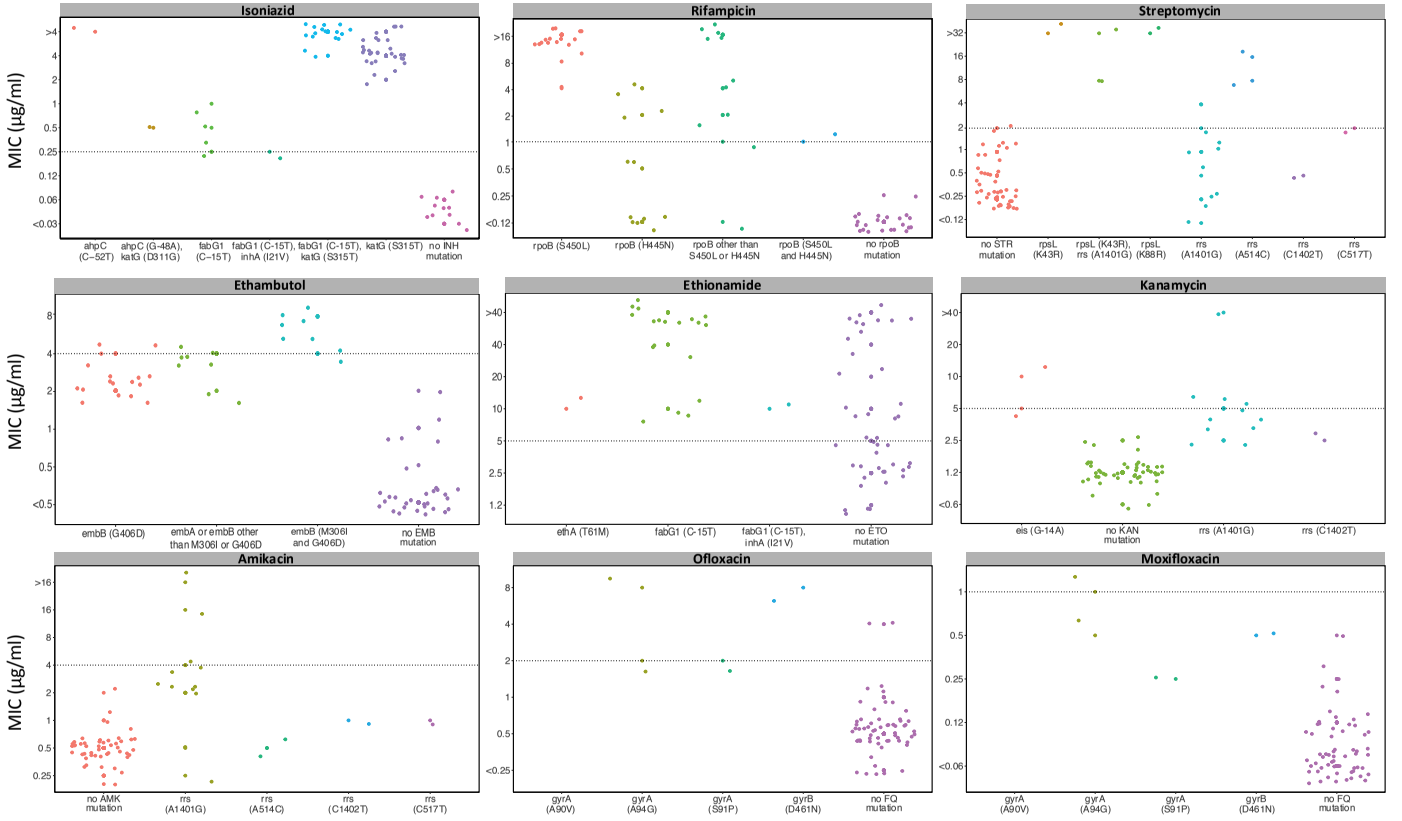


Figure 3.1. Drug resistance mutations with corresponding minimum inhibitory concentrations for nine antituberculosis drugs. Each dot represents an isolate and is coloured by mutation. *AhpC* C-52T, G-48A, *fabG1* C-15T, and *eis* G-14A are mutations in the promotor area of the respective gene. The y-axes show the minimum inhibitory concentrations in µg/ml. NOTE. Dots may be shown in-between tested MICs to increase readability.

according to MGIT, and 11 were MGIT-resistant, but had no STR resistance-conferring mutations. For EMB, we observed 34/69 (49%) discrepancies; all isolates without an EMB resistance-conferring mutation were susceptible, but 34 isolates with an EMB resistance mutation were also susceptible. Only isolates with the combination of two *embB* mutations (M306I and G406D) were MGIT EMB-resistant, and this was only the case for three out of nine isolates with this combined mutation.

We next compared WGS with the MycoTB assay, which yields MICs (Figure 3.1 and Table 3.1). Different resistance mutations were associated with different MIC ranges. For INH, the *katG* S315T mutation was associated with clearly elevated MICs, as was the *ahpC* C52T promotor mutation although it only occurred in one isolate. Other mutations in *ahpC*, *fabG1*, and *inhA* were associated with MICs around or just above the critical concentration. For RIF, the *rpoB* S450L mutation was associated with elevated MICs, but other rpoB mutations were associated with a range of MICs including the critical concentration. Mutations in *rpsL* were all associated with MICs above the critical concentration for STR and the same applies to the *rrs* A514C mutation. However, other *rrs* mutations were associated with MICs around or below the critical concentration for STR. All found mutations in *embA* or *embB* were associated with elevated EMB MICs; their ranges included the critical concentration.

All isolates with resistance-conferring mutations to ETO had elevated MICs. The *fabG1* C15T promotor mutation was associated with a bimodal MIC distribution for ETO; some isolates had an MIC of twice the critical concentration; most others with the mutation had an eight times higher MIC than the critical concentration. For AMK and KAN, all observed mutations were associated with an MIC range below or including the critical concentration. OFL or MXF resistance mutations were uncommon, and these mutations were associated with an MIC above the critical concentration in only two isolates, and only for OFL. For STR and OFL, but especially for ETO, we observed isolates with no mutations conferring resistance to the respective drug, but with MICs above the critical concentration.

Discussion

In this study, we demonstrate that MIC determination using SensiTitre MycoTB is an excellent alternative to conventional DST methods; it is relatively rapid, straightforward, and it provides quantitative data on susceptibility to first- and second-line drugs, thus facilitating therapeutic decision-making and therapeutic drug monitoring to optimize regimen efficacy and minimize toxicity[20]. Whole genome sequencing has been shown to be a sensitive, accurate, rapid, and financially feasible method for *M.*

tuberculosis drug susceptibility testing[21]. Here we show that WGS has additional added value in terms of predicting the level of resistance to different drugs. For key anti-tuberculosis drugs including rifampicin, isoniazid, ethambutol and fluoroquinolones, distinct genomic mutations are associated with particular MIC ranges, some including the critical concentration.

The MycoTB microdilution method is a commercially available method for MIC testing in *M. tuberculosis* and the first for second-line drugs. Previous studies have tested its performance compared to currently available methods for drug susceptibility testing, such as the BACTEC MGIT 960[22], the indirect agar proportion method[8,9,23], the Löwenstein-Jensen proportion method[24], genotypic tests[25], or a combination of these[17], and it has proven a rapid, convenient, quantitative, and accurate method for testing both first- and second-line antituberculosis drug susceptibility. We found discrepancies with WHO-endorsed phenotypic tests for first line drugs, especially STR and EMB, but phenotypic DST has been proven to be difficult for these drugs and the value of these tests as 'gold standard' is questionable. Compared to previous studies comparing MycoTB with MGIT, the MycoTB in the present study performed slightly better for INH and RIF, comparable for EMB, and slightly worse for STR[17,22]. Furthermore, the microwell plate format without the need for equipment will allow its use in resource-pour settings, where it is most needed, provided that culture-based DST is possible, and proper biosafety measures are taken. Moreover, the visual reading of the plates requires some operator experience.

Whole genome sequencing is revolutionizing drug susceptibility testing in tuberculosis and has great potential in public health interventions. However, the various platforms used, as well as the shortcomings of critical concentrations for phenotypic DST hinder the comparison of genotypic and phenotypic DST, hence the validation of WGS-based susceptibility testing[26]. The current critical concentrations have a limited evidence base and stem largely from observations on wild type MIC distributions, not from clinical or pharmacokinetics/pharmacodynamics studies[27]. Data from several studies indicate that these critical concentrations need to be revised because they either bisect the wild-type distribution, or are substantially higher than the MICs of wild-type organisms resulting in potential false reporting of susceptibility[28]. Indeed, in the current study, we observed that isolates with mutations associated with decreased susceptibility to RIF, STR, and EMB were reported susceptible according to MGIT, and MICs for these drugs were often around the critical concentration. These data support the notion of the European Committee on Antimicrobial Susceptibility Testing (EUCAST) that critical concentrations should be defined by combining MIC distributions, preferably combined with clinical outcomes and pharmacokinetics/pharmacodynamics data[28,29]. Even though WGS has already been successfully implemented in routine

diagnostic practice in some settings and has been shown to achieve generally high agreement with phenotypic first-line DST[21,30,31], MIC testing may help in more accurately assessing the performance of WGS for drug resistance detection and the role of this method in TB laboratory diagnosis.

The discrepancies between MGIT and MICs observed in our study support the current dogma that inconsistencies between phenotypic and genotypic DST found in important studies investigating the use of WGS for predicting *M. tuberculosis* drug susceptibility are partly attributed to shortcomings with currently endorsed methods of phenotypic DST, which essentially provide only qualitative results (sensitive or resistant), especially where mutations conferring low-level drug resistance are involved[32]. However, it is important to appreciate that unknown resistance mechanisms, inadequate limits of detection or artefacts of sequencing, random errors, and false associations between genotype and phenotype due to epistatic interactions could all play a role[33–35]. The clinical impact of these discrepancies, and the effect of treatment regimen based on different DST methods are not clear yet, and clinical validation of the influence of MICs of single drugs on treatment outcome is warranted. The analysis of *embB* mutations has long been considered futile because they did not match well with phenotypic DST results[36], our data however show that these mutations do have an effect on MICs thus potentially influence treatment efficacy[37,38] as has been shown previously in a mouse model of aerogenic tuberculosis[39].

Similar to what has been shown in a recent study by Heyckendorf *et al.*[32], we observed that WGS did not miss phenotypically confirmed resistances to first-line drugs, except for STR. We discovered that the MIC-range for isolates without STR resistance-conferring mutations was wide and included the critical concentration. Unknown mutations in these isolates could have caused resistance. Critical concentration artefacts could also explain the discrepancies; 16 isolates were MGIT-resistant for STR, but had an MIC below the critical concentration. In addition, WGS predicted resistance in a number of phenotypically susceptible isolates. Especially *rpoB* H445N, *embB* G406D and M306I, as well as the *rrs* mutations did not correlate well with phenotypic resistance, in line with previous findings[19,26,40–42]. We found that these mutations increased the MIC only slightly, or were associated with wide MIC-ranges. However, previous studies have shown that these mutations may still be clinically relevant[37,39,41,43,44].

This is the first study to investigate drug resistance in *M. tuberculosis* isolates from the greater Craiova area, Romania. In this selection of phenotypically drug-resistant isolates we observed low frequencies of resistance to second-line drugs, possibly as a result of limited availability of second-line antituberculosis drugs[7]. Surprising was the observation that all-but-one of the isolates belonged to the Euro-American lineage,

because there has been an extremely large cluster of MDR/XDR-TB *M. tuberculosis* strains in the EU, especially in the eastern part, which is significantly related to the spread of one strain or clone of the Beijing genotype[45].

The relatively small number of isolates does not permit drawing solid conclusions on the effect of most drug resistance mutations on drug susceptibility and the interrelatedness of different mutations, or their combined effects on the level of resistance. However, the findings we present are an important contribution to the field because of a lack of data correlating *M. tuberculosis* genotype and MICs[46]. In addition, the phylogeny showed one clade of genetically closely related isolates. This could have affected the MIC distribution of the drugs these isolates are resistant to. However, the drug resistance mutations found in these isolates differed to an extent that makes it unlikely that these isolates represent the same strain. Also, we could not associate drug resistance mutations with the corresponding MICs for two strains that failed to grow on the MycoTB microtiter plate and six that were contaminated. Lastly, pyrazinamide is not included in the MycoTB plate, and we could not assess the MICs related to pyrazinamide resistance-associated mutations, which were discovered in 23 (32%) isolates.

In summary, we have shown that, especially for RIF, STR, and EMB, MICs near the critical concentration are common. Consequently, phenotypic DST based on critical concentration testing, e.g. the MGIT method, may provide inaccurate results, possibly leading to suboptimal treatment regimens. We compared WGS-predicted drug resistance mutations directly with MICs and found that different mutations lead to different levels of resistance; knowing the underlying mutations can guide clinical decision-making and facilitate therapeutic drug monitoring, ultimately leading to better treatment outcome.

Author Contributions

A.L.R., M.I., R.v.C., and J.v.I. designed the study, A.L.R., A.F., L.C., C.E., N.A., D.N., and V.G. performed the isolation, cultivation, and sequencing tasks, C.R., L.C., N.A., R.v.C., and J.v.I. analysed the data, C.R., A.L.R., L.C., R.v.C., and J.v.I. wrote the manuscript, and all authors read and approved the final version of the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-27962-5.

Competing Interests

The authors declare no competing interests.

Supplementary figures

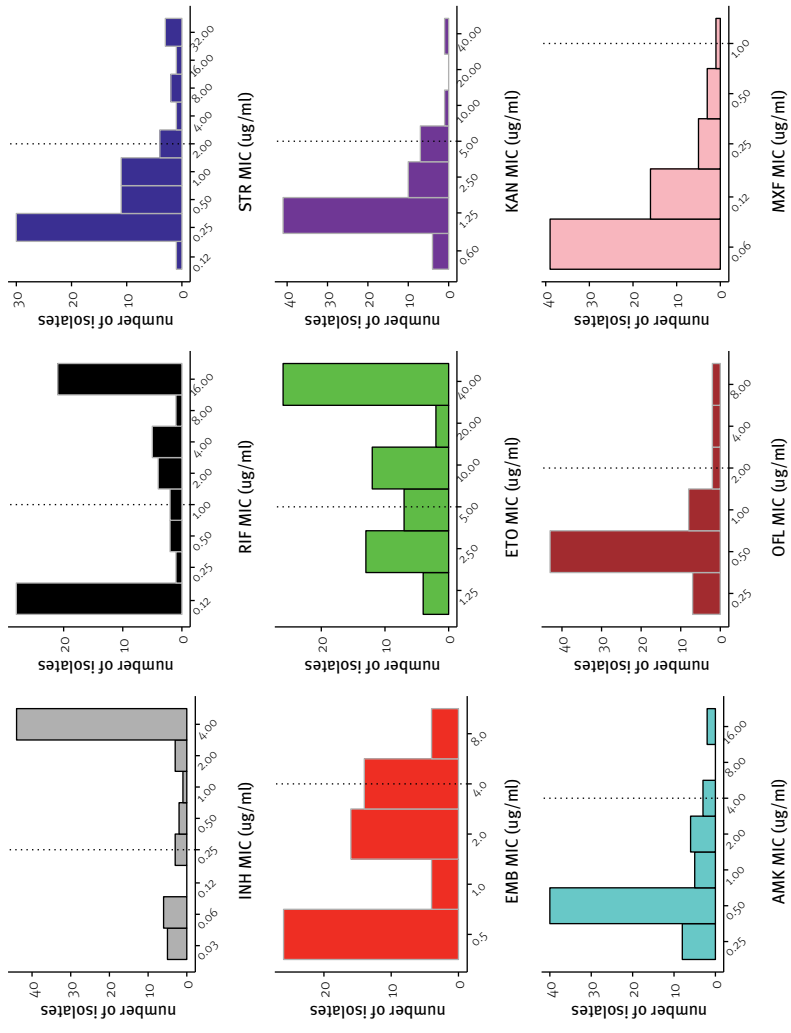


Figure S3.1. Overview of minimum inhibitory concentrations for nine antituberculous drugs. Minimum inhibitory concentrations were determined using the MycoTB microdilution method for 64 *M. tuberculosis* isolates.

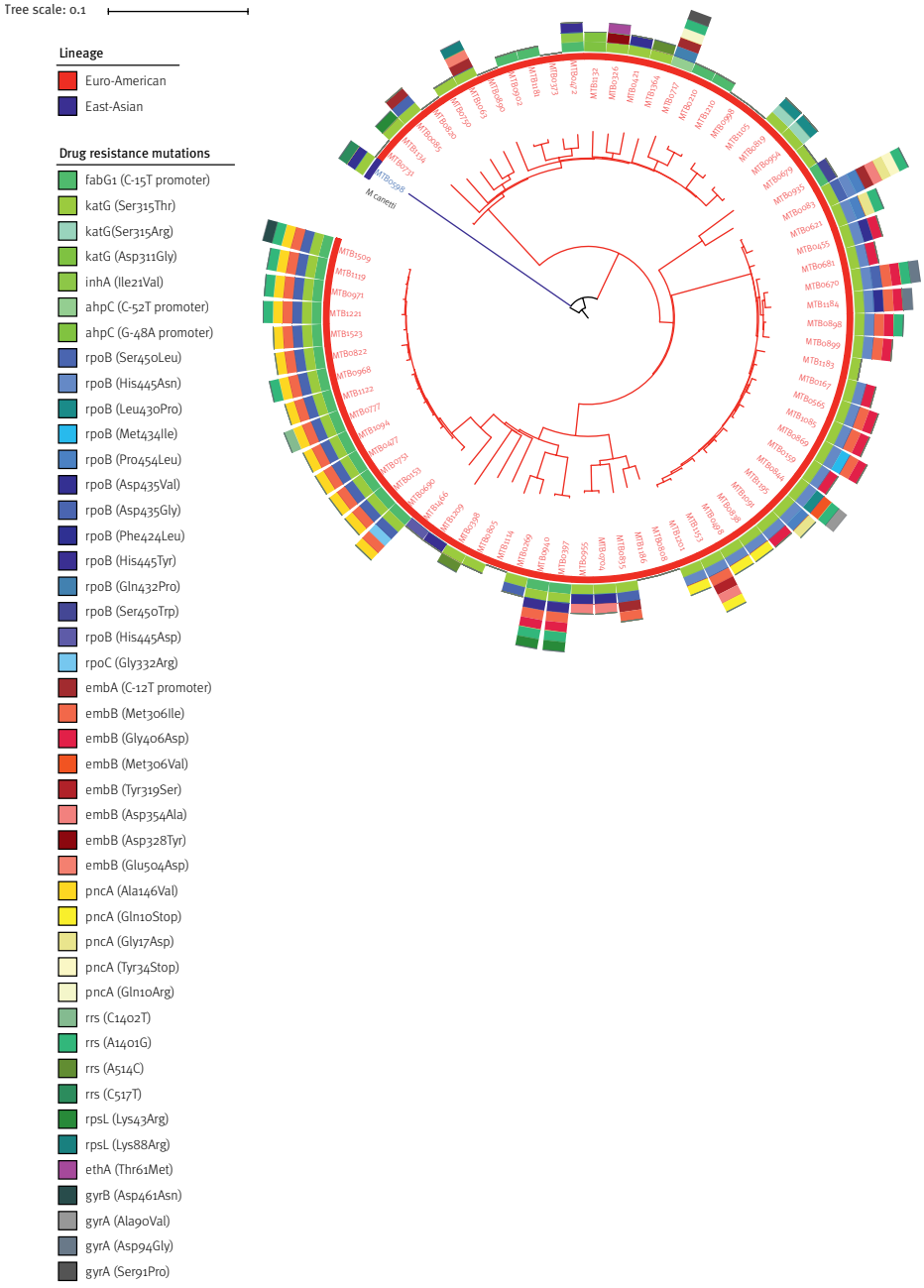


Figure S3.2. Phylogenetic tree showing the genetic relatedness of 72 *M. tuberculosis* strains isolated from 72 Romanian patients.

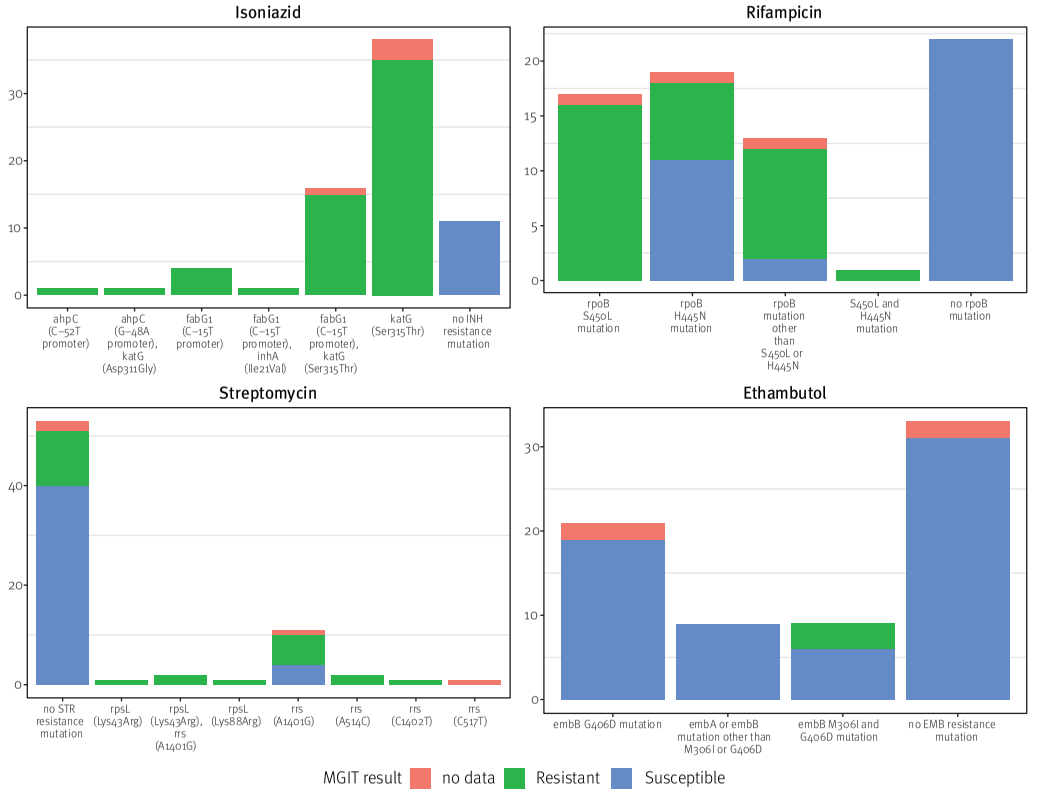


Figure S3.3. Comparison of whole genome sequencing and Mycobacterium Growth Indicator Tube to determine drug susceptibility to first-line anti-tuberculosis drugs.

Supplementary tables

Table S3.1. Description of sequencing quality control parameters and statistics

Study number	Clean reads	Q20%	Total bases	Mean coverage	Median coverage	Percentage bases >1
MTB0083	5010168	98,8	729245204	155,8	160	99,6
MTB0085	9244564	98,9	1331280683	292,3	299	99,4
MTB0153	4980076	98,8	720076577	157,9	161	99,5
MTB0159	4326888	98,8	628038812	138,1	142	99,6
MTB0167	5439392	98,9	792319455	174,2	178	99,6
MTB0210	5880320	98,9	852347773	187,4	192	99,5
MTB0269	4613254	98,8	671851858	147,7	151	99,7
MTB0326	5072238	98,8	730714611	160,2	164	99,5
MTB0373	4397414	98,8	639876834	139,8	143	99,5
MTB0397	2668762	98,8	388958930	82,6	85	99,8
MTB0398	3282760	98,7	479249813	103,1	105	99,5
MTB0421	5498680	98,8	797711420	174,6	179	99,5
MTB0455	2581286	98,8	377936416	79,3	78	99,5
MTB0472	6458404	98,8	937346421	204,5	210	99,5
MTB0477	5466848	98,8	793482681	173,6	178	99,5
MTB0498	3500310	98,7	509420826	108,6	112	99,5
MTB0565	6226184	98,8	907091707	198,6	204	99,5
MTB0598	5334456	98,7	776163802	169,7	173	99,3
MTB0621	1315430	98,7	192150033	40,4	41	99,4
MTB0670	1723318	98,7	252717866	51,7	53	99,5
MTB0679	5047768	98,8	735251932	160,7	164	99,7
MTB0681	2592134	98,3	374789416	81,0	82	99,5
MTB0690	2788926	98,7	407670628	88,0	90	99,5
MTB0704	4512874	98,8	655016172	143,0	146	99,4
MTB0717	5180780	98,8	753070029	164,8	169	99,5
MTB0731	6910880	98,8	995135438	218,9	225	99,7
MTB0750	16678804	98,8	2289895133	501,3	500	99,6
MTB0751	5024722	98,8	725893624	159,6	160	99,6
MTB0777	6829294	98,8	984893721	216,8	222	99,6
MTB0805	6349612	98,9	920541812	202,5	208	99,8
MTB0808	13982680	98,9	1992277055	438,3	450	99,6
MTB0819	7285102	98,8	1045110908	228,2	233	99,3
MTB0820	6437616	98,8	928788036	203,9	209	99,6
MTB0822	10553814	98,8	1498369699	329,4	335	99,6
MTB0835	13273726	98,9	1867111990	410,9	422	99,5

Table S3.1. Continued

Study number	Clean reads	Q20%	Total bases	Mean coverage	Median coverage	Percentage bases >1
MTB0838	11449646	98,9	1631527546	358,9	368	99,6
MTB0844	7042508	98,8	1014627854	223,5	229	99,6
MTB0869	9514364	98,8	1348716932	296,8	304	99,6
MTB0890	9986614	98,8	1426980119	312,9	321	99,5
MTB0898	2673806	98,8	389093509	78,8	80	99,5
MTB0899	3748466	98,8	543659909	110,8	111	99,6
MTB0902	5105332	98,8	738012399	161,6	166	99,5
MTB0935	4166006	98,8	600850392	131,6	134	99,9
MTB0940	6345710	98,9	912105253	200,3	205	99,8
MTB0954	5182548	98,9	749833435	164,1	169	99,3
MTB0955	6342280	98,9	912264427	200,4	206	99,4
MTB0968	5597674	98,8	808331240	177,3	182	99,6
MTB0971	4810290	98,8	697511432	153,1	158	99,5
MTB0998	6125292	98,8	882233413	193,8	200	99,5
MTB1063	3131132	98,8	452357038	99,1	102	99,5
MTB1085	4055568	98,8	586012462	127,2	131	99,5
MTB1091	4496052	98,8	651171048	143,0	147	99,5
MTB1094	5037192	98,8	726608727	159,5	164	99,5
MTB1105	2766362	98,8	403334631	86,0	88	99,5
MTB1114	7021708	98,8	1005915329	221,2	227	99,7
MTB1119	4133620	98,7	598668119	131,1	134	99,5
MTB1122	5510106	98,8	800687586	175,7	179	99,5
MTB1132	2041418	98,7	297936222	64,9	66	99,2
MTB1134	2636480	98,8	384682814	84,1	86	99,3
MTB1153	2592910	98,8	377531940	82,8	85	99,5
MTB1181	3687154	98,8	536107686	117,5	121	99,5
MTB1183	4231256	98,8	614868223	135,1	138	99,5
MTB1184	1520090	98,8	222159082	47,7	49	99,5
MTB1186	2331846	98,7	339340987	74,3	76	99,5
MTB1195	5434990	98,7	786889583	172,7	177	99,6
MTB1201	3044342	98,7	439656581	96,5	99	99,5
MTB1209	3420894	98,7	494659829	108,5	111	99,5
MTB1221	2829856	98,8	411306820	89,8	91	99,5
MTB1364	3977738	98,7	576100368	125,7	128	99,5
MTB1466	2388624	98,7	347922516	75,6	77	99,6
MTB1509	3914664	98,6	568644763	124,2	126	99,5
MTB1523	2993102	98,8	435834604	95,7	98	99,5

Table S3.2. Discrepant drug susceptibility test results according to whole genome sequencing and Mycobacterium Growth Indicator Tube.

				Sequencing statistics		
Isolate	Discrepant WGS result	Discrepant MGIT result	LJ result	Underlying mutation	Sequencing coverage depth (no. of reads)	% of reads supporting the variant
MTB0159	RIF-resistant	RIF-susceptible	RIF-susceptible	rpoB (H445N)	147	100%
MTB0498	RIF-resistant	RIF-susceptible	RIF-susceptible	rpoB (H445N)	115	96.5%
MTB0598	RIF-resistant	RIF-susceptible	Rif-resistant	rpoB (D435V)	161	97.5%
MTB0621	RIF-resistant	RIF-susceptible	Rif-resistant	rpoB (H445N)	37	100%
				rpoB (P454L)	48	100%
MTB0681	RIF-resistant	RIF-susceptible	RIF-susceptible	rpoB (H445N)	123	97.6%
MTB0819	RIF-resistant	RIF-susceptible	RIF-susceptible	rpoB (L430P)	234	99.6%
MTB0838	RIF-resistant	RIF-susceptible	RIF-susceptible	rpoB (H445N)	383	98.4%
MTB0898	RIF-resistant	RIF-susceptible	RIF-susceptible	rpoB (H445N)	78	97.4%
MTB0899	RIF-resistant	RIF-susceptible	RIF-susceptible	rpoB (H445N)	122	99.2%
MTB1085	RIF-resistant	RIF-susceptible	RIF-susceptible	rpoB (H445N)	130	99.2%
MTB1091	RIF-resistant	RIF-susceptible	RIF-susceptible	rpoB (H445N)	136	99.3%
MTB1153	RIF-resistant	RIF-susceptible	RIF-susceptible	rpoB (H445N)	79	100%
MTB1195	RIF-resistant	RIF-susceptible	RIF-susceptible	rpoB (H445N)	183	99.5%

Table S3.3. Concordance between Mycobacterium Growth Indicator Tube and MycoTB

			MycoTB			
			MIC above breakpoint (resistant)	MIC below breakpoint (susceptible)	% Categorical agreement	Cohen's kappa (p-value)
MGIT	INH	Resistant	48	3	93.5	0.794 (p<0.001)
		Susceptible	1	10		
	RIF	Resistant	29	2	95.2	0.905 (p<0.001)
		Susceptible	1	31		
	STR	Resistant	7	16	71.0	0.289 (p=0.006)
		Susceptible	2	37		
	EMB	Resistant	1	2	92.3	0.246 (p=0.045)
		Susceptible	3	59		

References

1 WHO. Global tuberculosis report 2016.

2 WHO. *Multidrug and extensively drug-resistant TB (M/XDR-TB): 2010 global report on surveillance and response*, http://apps.who.int/iris/bitstream/10665/44286/1/ 9789241599191_eng.pdf

3 Horne, D. J. *et al.* Diagnostic accuracy and reproducibility of WHO-endorsed phenotypic drug susceptibility testing methods for first-line and second-line antituberculosis drugs. *J Clin Microbiol* **51**, 393-401, doi: 10.1128/JCM.02724-12 (2013).

4 Schon, T. *et al. Mycobacterium tuberculosis* drug-resistance testing: challenges, recent developments and perspectives. *Clin Microbiol Infect* **23**, 154-160, doi:10.1016/j.cmi.2016.10.022 (2017).

5 Steingart, K. R. *et al.* Xpert(R) MTB/RIF assay for pulmonary tuberculosis and rifampicin resistance in adults. *Cochrane Database Syst Rev*, CD009593, doi:10.1002/14651858.CD009593.pub3 (2014).

6 Theron, G. *et al.* The diagnostic accuracy of the GenoType((R)) MTBDRsl assay for the detection of resistance to second-line anti-tuberculosis drugs. *Cochrane Database Syst Rev*, CD010705, doi:10.1002/14651858.CD010705.pub2 (2014).

7 WHO. Review of the National Tuberculosis Programme in Romania, 10–21 March 2014. (2015).

8 Abuali, M. M., Katariwala, R. & LaBombardi, V. J. A comparison of the Sensititre(R) MYCOTB panel and the agar proportion method for the susceptibility testing of *Mycobacterium tuberculosis*. *Eur J Clin Microbiol Infect Dis* **31**, 835-839, doi:10.1007/s10096-011-1382-z (2012).

9 Hall, L. *et al.* Evaluation of the Sensititre MycoTB plate for susceptibility testing of the *Mycobacterium tuberculosis* complex against first- and second-line agents. *J Clin Microbiol* **50**, 3732-3734, doi:10.1128/JCM.02048-12 (2012).

10 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

11 Coll, F. *et al.* Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Med* **7**, 51, doi:10.1186/s13073-015-0164-0 (2015).

12 Coll, F. *et al.* PolyTB: a genomic variation map for *Mycobacterium tuberculosis*. *Tuberculosis (Edinb)* **94**, 346-354, doi:10.1016/j.tube.2014.02.005 (2014).

13 Feuerriegel, S. *et al.* PhyResSE: a Web Tool Delineating *Mycobacterium tuberculosis* Antibiotic Resistance and Lineage from Whole-Genome Sequencing Data. *J Clin Microbiol* **53**, 1908-1914, doi:10.1128/JCM.00025-15 (2015).

14 Andre, E. *et al.* Consensus numbering system for the rifampicin resistance-associated rpoB gene mutations in pathogenic mycobacteria. *Clin Microbiol Infect* **23**, 167-172, doi:10.1016/j.cmi.2016.09.006 (2017).

15 Barrick, J. E. Identifying structural variation in haploid microbial genomes from short-read resequencing data using breseq. *BMC Genomics* (2014).

16 Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**, 307-321, doi:10.1093/sysbio/syq010 (2010).

17 Banu, S. *et al.* Discordance across several methods for drug susceptibility testing of drug-resistant *Mycobacterium tuberculosis* isolates in a single laboratory. *J Clin Microbiol* **52**, 156-163, doi:10.1128/JCM.02378-13 (2014).

18 Reeves, A. Z., Campbell, P. J., Willby, M. J. & Posey, J. E. Disparities in capreomycin resistance levels associated with the *rrs* A1401G mutation in clinical isolates of *Mycobacterium tuberculosis*. *Antimicrob Agents Chemother* **59**, 444-449, doi:10.1128/AAC.04438-14 (2015).

19 Miotto, P. *et al.* A standardised method for interpreting the association between mutations and phenotypic drug resistance in *Mycobacterium tuberculosis*. *Eur Respir J* **50**, doi:10.1183/13993003.01354-2017 (2017).

20 Heysell, S. K., Moore, J. L., Peloquin, C. A., Ashkin, D. & Houpt, E. R. Outcomes and use of therapeutic drug monitoring in multidrug-resistant tuberculosis patients treated in virginia, 2009-2014. *Tuberc Respir Dis (Seoul)* **78**, 78-84, doi:10.4046/trd.2015.78.2.78 (2015).

21 Pankhurst, L. J. *et al.* Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: a prospective study. *The Lancet Respiratory Medicine* **4**, 49-58, doi:10.1016/s2213-2600(15)00466-x (2016).

22 Heysell, S. K. *et al.* Sensititre MycoTB plate compared to Bactec MGIT 960 for first- and second-line antituberculosis drug susceptibility testing in Tanzania: a call to operationalize MICs. *Antimicrob Agents Chemother* **59**, 7104-7108, doi:10.1128/AAC.01117-15 (2015).

23 Lee, J. *et al.* Sensititre MYCOTB MIC plate for testing *Mycobacterium tuberculosis* susceptibility to first- and second-line drugs. *Antimicrob Agents Chemother* **58**, 11-18, doi:10.1128/AAC.01209-13 (2014).

24 Yu, X. *et al.* Sensititre(R) MYCOTB MIC plate for drug susceptibility testing of *Mycobacterium tuberculosis* complex isolates. *Int J Tuberc Lung Dis* **20**, 329-334, doi:10.5588/ijtld.15.0573 (2016).

25 Foongladda, S. *et al.* Comparison of TaqMan((R)) Array Card and MYCOTB(TM) with conventional phenotypic susceptibility testing in MDR-TB. *Int J Tuberc Lung Dis* **20**, 1105-1112, doi:10.5588/ijtld.15.0896 (2016).

26 Dominguez, J. *et al.* Clinical implications of molecular drug resistance testing for *Mycobacterium tuberculosis*: a TBNET/RESIST-TB consensus statement. *Int J Tuberc Lung Dis* **20**, 24-42, doi:10.5588/ijtld.15.0221 (2016).

27 Gumbo, T., Angulo-Barturen, I. & Ferrer-Bazaga, S. Pharmacokinetic-pharmacodynamic and dose-response relationships of antituberculosis drugs: recommendations and standards for industry and academia. *J Infect Dis* **211 Suppl 3**, S96-S106, doi:10.1093/infdis/jiu610 (2015).

28 Angeby, K., Jureen, P., Kahlmeter, G., Hoffner, S. E. & Schon, T. Challenging a dogma: antimicrobial susceptibility testing breakpoints for *Mycobacterium tuberculosis*. *Bull World Health Organ* **90**, 693-698, doi:10.2471/BLT.11.096644 (2012).

29 Kahlmeter, G. The 2014 Garrod Lecture: EUCAST - are we heading towards international agreement? *J Antimicrob Chemother* **70**, 2427-2439, doi:10.1093/jac/dkv145 (2015).

30 Quan, T. P. *et al.* Evaluation of Whole-Genome Sequencing for Mycobacterial Species Identification and Drug Susceptibility Testing in a Clinical Setting: a Large-Scale Prospective Assessment of Performance against Line Probe Assays and Phenotyping. *J Clin Microbiol* **56**, doi:10.1128/JCM.01480-17 (2018).

31 Shea, J. *et al.* Comprehensive Whole-Genome Sequencing and Reporting of Drug Resistance Profiles on Clinical Cases of *Mycobacterium tuberculosis* in New York State. *J Clin Microbiol* **55**, 1871-1882, doi:10.1128/JCM.00298-17 (2017).

32 Heyckendorf, J. *et al.* What is resistance? Impact of phenotypic versus molecular drug resistance testing on multi- and extensively drug-resistant tuberculosis therapy. *Antimicrob Agents Chemother*, doi:10.1128/AAC.01550-17 (2017).

33 Bradley, P. *et al.* Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. *Nat Commun* **6**, 10063, doi:10.1038/ncomms10063 (2015).

34 Fenner, L. *et al.* Effect of mutation and genetic background on drug resistance in *Mycobacterium tuberculosis*. *Antimicrob Agents Chemother* **56**, 3047-3053, doi:10.1128/AAC.06460-11 (2012).

35 Koser, C. U., Feuerriegel, S., Summers, D. K., Archer, J. A. & Niemann, S. Importance of the genetic diversity within the *Mycobacterium tuberculosis* complex for the development of novel antibiotics and diagnostic tests of drug resistance. *Antimicrob Agents Chemother* **56**, 6080-6087, doi:10.1128/AAC.01641-12 (2012).

36 Cheng, S., Cui, Z., Li, Y. & Hu, Z. Diagnostic accuracy of a molecular drug susceptibility testing method for the antituberculosis drug ethambutol: a systematic review and meta-analysis. *J Clin Microbiol* **52**, 2913-2924, doi:10.1128/JCM.00560-14 (2014).

37 Williamson, D. A. *et al.* Clinical failures associated with *rpoB* mutations in phenotypically occult multi-drug-resistant *Mycobacterium tuberculosis*. *Int J Tuberc Lung Dis* **16**, 216-220, doi:10.5588/ijtld.11.0178 (2012).

38 Ho, J., Jelfs, P. & Sintchencko, V. Phenotypically occult multidrug-resistant *Mycobacterium tuberculosis*: dilemmas in diagnosis and treatment. *J Antimicrob Chemother* **68**, 2915-2920, doi:10.1093/jac/dkt284 (2013).

39 Plinke, C., Walter, K., Aly, S., Ehlers, S. & Niemann, S. *Mycobacterium tuberculosis embB* codon 306 mutations confer moderately increased resistance to ethambutol *in vitro* and *in vivo*. *Antimicrob Agents Chemother* **55**, 2891-2896, doi:10.1128/AAC.00007-10 (2011).

40 Witney, A. A. *et al.* Clinical use of whole genome sequencing for *Mycobacterium tuberculosis*. *BMC Med* **14**, 46, doi:10.1186/s12916-016-0598-2 (2016).

41 Jamieson, F. B. *et al.* Profiling of *rpoB* mutations and MICs for rifampin and rifabutin in *Mycobacterium tuberculosis*. *J Clin Microbiol* **52**, 2157-2162, doi:10.1128/JCM.00691-14 (2014).

42 Naidoo, C. C. & Pillay, M. Fitness-compensatory mutations facilitate the spread of drug-resistant F15/LAM4/KZN and F28 *Mycobacterium tuberculosis* strains in KwaZulu-Natal, South Africa. *Journal of Genetics* **96**, 599-612, doi:10.1007/s12041-017-0805-8 (2017).

43 Schon, T. *et al.* Rifampicin-resistant and rifabutin-susceptible *Mycobacterium tuberculosis* strains: a breakpoint artefact? *J Antimicrob Chemother* **68**, 2074-2077, doi:10.1093/jac/dkt150 (2013).

44 Rigouts, L. *et al.* Rifampin resistance missed in automated liquid culture system for *Mycobacterium tuberculosis* isolates with specific *rpoB* mutations. *J Clin Microbiol* **51**, 2641-2645, doi:10.1128/JCM.02741-12 (2013).

45 De Beer, J. L. *et al.* Molecular surveillance of multi- and extensively drug-resistant tuberculosis transmission in the European Union from 2003 to 2011. *Euro Surveill* **19** (2014).

46 Burke, R. M., Coronel, J. & Moore, D. Minimum inhibitory concentration distributions for first- and second-line antimicrobials against *Mycobacterium tuberculosis*. *J Med Microbiol* **66**, 1023-1026, doi:10.1099/jmm.0.000534 (2017).

4

Diabetes is associated with genotypically drug-resistant tuberculosis

Carolien Ruesen, Lidya Chaidir, Cesar Ugarte-Gil, Jakko van Ingen,
Julia A. Critchley, Philip C. Hill, Rovina Ruslami, Prayudi Santoso, Martijn A. Huynen,
Hazel M. Dockrell, David A.J. Moore, Bachti Alisjahbana, Reinout van Crevel.

Eur Resp J. To be adapted for resubmission.

Abstract

Background

Many tuberculosis (TB) patients with diabetes have poor treatment outcomes, but it is unclear whether this results from genotypic drug resistance of the infecting mycobacteria.

Methods

In this observational study we used whole genome sequencing to examine 1,365 known drug resistance mutations in 896 *Mycobacterium tuberculosis* isolates from TB patients from Indonesia and Peru with (n=159) and without (n=737) diabetes, 19% and 24% of whom respectively had been previously treated for TB. We used multilevel logistic regression to examine the relationship between diabetes and resistance mutations.

Results

In Indonesia, drug resistance mutations were found in *M. tuberculosis* isolates of 21 / 115 (18%) patients with and 43 / 292 (15%) patients without diabetes, respectively. In Peru, mutations were found in 18 / 44 (41%) patients with, compared to 149 / 445 (33%) without diabetes. Diabetes was independently associated with genotypic drug resistance (odds ratio (OR) = 1.69; 95% confidence interval (CI) = 1.04-2.77), in particular to rifampicin (OR = 2.52 (95% CI 1.19-5.34)). The association was stronger for those not previously treated for TB. Mutations in *Rv1482c-fabG1*, conferring resistance to isoniazid and ethionamide, and *gyrA*, conferring fluoroquinolone resistance, were overrepresented in isolates from Peruvian patients with diabetes ($p<0.01$ and $p<0.05$, respectively).

Conclusions

Diabetes is associated with genotypic drug resistance in *M. tuberculosis*, and this is not explained by prior TB treatment. Our findings support the importance of routine drug-susceptibility testing for TB patients with diabetes, and highlight the need for further study to understand the mechanism underlying the association.

Introduction

The convergence of the tuberculosis (TB) and type 2 diabetes epidemics[1] poses a great threat to global TB control. People with diabetes are three times more likely to develop active TB than people without diabetes, and diabetes is associated with increased death, TB treatment failure and recurrent TB[2-5]. A higher prevalence of drug-resistant TB among those with diabetes may be one contributing factor to their adverse clinical outcomes during and after TB treatment. Two recent meta-analyses examining the association between diabetes and multidrug-resistant TB (MDR-TB), including more than twenty-two thousand individuals with TB, showed that diabetes is an independent risk factor for MDR-TB[3,6]. Drug-resistant TB is associated with increased morbidity and mortality and has emerged as a major threat to TB control worldwide[7]. However, the genetic characteristics of resistant *M. tuberculosis* strains causing disease in diabetes patients remain to be discovered.

In the current study we used whole genome sequencing to investigate genetic differences of *M. tuberculosis* isolates causing disease in two cohorts of TB patients with and without diabetes, focusing on the presence of drug resistance-associated mutations. As opposed to phenotypic drug susceptibility testing (DST), whole genome sequencing has the potential to unlock valuable information on the specific mutations underlying resistance, transmission clusters and the phylogenetic background of drug-resistant *M. tuberculosis* strains[8], thereby increasing our understanding of the molecular basis of resistance of *M. tuberculosis* in these patients.

Methods

Subjects and isolates

We used *M. tuberculosis* isolates from two established TB cohorts of Indonesian and Peruvian patients who were screened for diabetes. Both cohorts were part of TANDEM, a multi-centre study focussing on the relationship between TB and diabetes[9-11]. All participants underwent laboratory glycated hemoglobin (HbA1c) testing (using the high-performance liquid chromatography method), as recommended by WHO for diabetes screening[12], regardless of their previous diabetes status. Further details on TB and diabetes case definitions are published elsewhere[10,13] and are freely available in online appendices[14]. In Peru we selected all available *M. tuberculosis* isolates; in Indonesia we selected all available isolates from TB patients with diabetes plus a subset of isolates from patients without diabetes from the same clinics, during the same time period, frequency-matched by age. A single isolate from each patient, obtained at time of diagnosis, was selected for sequencing. Ethical approval was

received from the Observational/Interventions Research Ethics Committee, London School of Hygiene and Tropical Medicine on 18 December 2013 (LSHTM ethics ref: 6449, LSHTM amendment no: A473) and institutional review boards in Indonesia and Peru.

Cultured *M. tuberculosis* isolates underwent whole genome sequencing as described in the Supplementary methods. We used TB Profiler version 0.3.8[15] to determine *M. tuberculosis* lineage and drug resistance, comparing identified single nucleotide polymorphisms (SNPs) and indels to a curated list of drug resistance mutations. All non-synonymous SNPs in *rpoA*, *rpoB* outside the rifampicin resistance-determining region (RRDR), *rpoC*, *ahpC* promoter region and *ubiA* were considered as mutations possibly compensating for the loss of fitness caused by drug resistance mutations[16-18].

A phylogeny was constructed using PhyML version 3.0[19]. Based on the phylogeny, we calculated the distance to the genetically closest other isolate (minimum pairwise distance) for isolates from TB patients with and without diabetes separately. We compared the distribution of these distances using the Mann-Whitney U test, for Indonesia and Peru separately.

Statistical analysis

We compared patient- and isolate-related characteristics of TB patients with and without diabetes, stratified by country (Indonesia or Peru). We assessed the association between diabetes and genotypic drug resistance using multilevel logistic regression, correcting for age, gender, HIV-infection, previous TB treatment, and *M. tuberculosis* lineage. We expanded the analysis by examining differences in resistance to individual drugs and combined mutations accounting for MDR-TB, again using multilevel multivariable logistic regression. To examine country-specific associations between diabetes and individual and multi-drug resistance, we used logistic regression stratified by country. We performed a sensitivity analysis to examine the effect of possible misclassification of patients with newly diagnosed diabetes on the association with any drug resistance. Furthermore, we investigated whether HbA1c was associated with drug resistance and also analysed the data after excluding all patients with HIV.

We investigated whether resistance mutations in certain genes occurred more frequently in TB patients with diabetes. Per resistance gene, we summed the number of isolates with at least one known resistance mutation in the respective gene and compared these between patients with and without diabetes, for Indonesia and Peru separately.

Lastly, we examined if a possible association between diabetes and drug resistance could be explained by differences in potential compensatory mutations, hypothesizing that resistant isolates from patients with diabetes, who are thought to have a lower

host immune defence against *M. tuberculosis*[20], would need fewer mutations compensating for loss of fitness associated with drug resistance. We determined the association between isoniazid and rifampicin resistance mutations and the presence of previously proposed accompanying compensatory mutations[16,18,21], stratified for diabetes comorbidity, for Indonesian and Peruvian isolates separately.

Results

Patient and isolate characteristics

Diabetes was more common among the Indonesian (28%) compared to the Peruvian patients (9%), but genotypic drug resistance was more often observed in isolates from Peruvian (34%) compared to Indonesian patients (16%, Table 4.1). In both countries, TB patients with diabetes were older, more often female, and slightly heavier than TB patients without diabetes. Only 21% of the Indonesian and 14% of the Peruvian diabetes patients had been previously treated for TB, compared to 27% and 22% of the non-diabetes patients, respectively (Table 4.1).

A phylogenetic tree, based on 56,654 variable common nucleotide positions among 896 *M. tuberculosis* isolates, showed that isolates from TB patients with diabetes were spread evenly across the phylogeny, i.e. they were not clustered (Figure 4.1). The *M. tuberculosis* lineage distribution for isolates from patients with diabetes was significantly different from those without (Chi-square = 11.559; p=0.009), with East-Asian strains being more common (30% vs. 20%) and Euro-American strains less common among the TB patients with diabetes (67% vs. 77%). However, this difference was no longer apparent after stratifying by country (Chi-square = 4.692; p=0.196 for Indonesia and Chi-square = 0.241; p=0.624 for Peru); lineage was indeed highly associated with country (Chi-square = 90.383; p<0.001). With respect to clonality: for isolates from Indonesian TB patients with diabetes, the median minimum pairwise distance was 161 SNPs; for patients without diabetes this amounted to 164 SNPs (p=0.867). For isolates from Peruvian TB patients with diabetes, the median minimum pairwise distance was 98 SNPs, compared to 67 SNPs in patients without diabetes (p=0.021, Supplementary figure S4.1).

Genotypic drug resistance in TB patients with and without diabetes

In Indonesia and Peru, respectively 18% and 39% of the patients with diabetes were infected with *M. tuberculosis* with at least one drug resistance-conferring mutation, compared to 15% and 20%, respectively, of the patients without diabetes (Fisher's exact p=0.454 and p=0.006 respectively; Table 4.1). In multilevel logistic regression, genotypic drug resistance was significantly associated with diabetes, both in

Table 4.1. Baseline characteristics of tuberculosis patients with and without diabetes mellitus type 2

	Indonesia		Peru	
	Diabetes (N=115)	No diabetes (N=292)	Diabetes (N=44)	No diabetes (N=445)
Male gender	52 (45%)	168 (58%)	25 (57%)	268 (60%)
Age (years) – median (IQR)	50 (45-58)	39 (33-48)	52 (42-59)	28 (22-39)
BMI	21 (18-25)	18 (16-20)	24 (22-28)	22 (20-24)
History of previous TB treatment	24 (21%)	79 (27%)	6 (14%)	96 (22%)
BCG scar	51 (65%)	108 (60%)	38 (86%)	342 (77%)
HIV-positive	0	0	1 (3%)	18 (5%)
M. tuberculosis lineage				
East-Asian	40 (35%)	90 (31%)	7 (16%)	59 (13%)
Euro-American	69 (60%)	178 (61%)	37 (84%)	386 (87%)
Indo-Oceanic	5 (4%)	24 (8%)	0	0
M. bovis	1 (1%)	0	0	0
Any resistance predicted by WGS	21 (18%)	44 (15%)	17 (39%)	88 (20%)
Treatment-naïve	13 (11%)	31 (11%)	14 (32%)	67 (15%)
Previously treated	8 (7%)	12 (4%)	3 (7%)	21 (5%)
MDR predicted by WGS	6 (5%)	8 (3%)	6 (14%)	38 (9%)

Data are presented as N (%) unless indicated otherwise.
Data are missing for BMI (Indonesia: n = 149), History of previous TB treatment (Indonesia: n = 3), BCG scar (Indonesia: n = 149), HIV-status (Indonesia: n = 1; Peru: n = 75).

univariable (odds ratio (OR) 1.6; 95% confidence interval (CI) 1.1-2.5), and multivariable analysis (OR 1.8; 95% CI 1.1-2.9). No other factors were independently associated with genotypic resistance (Table 4.2).

Combining both countries and accounting for the country effect using multilevel multivariable regression, diabetes was significantly associated with rifampicin resistance (OR 2.5; 95% CI 1.2-5.3), also among those not previously treated for TB (OR 3.1; 95% CI 1.2-8.3). In addition, we found a trend towards more fluoroquinolone resistance in TB patients with diabetes compared to those without, which reached significance for Peru (OR 6.69; 95% CI 1.37-32.68). Moreover, although this association did not reach statistical significance, the odds of MDR-TB was twice as high in patients with diabetes, compared to patients without (OR 2.09; 95% CI 0.92-4.77).

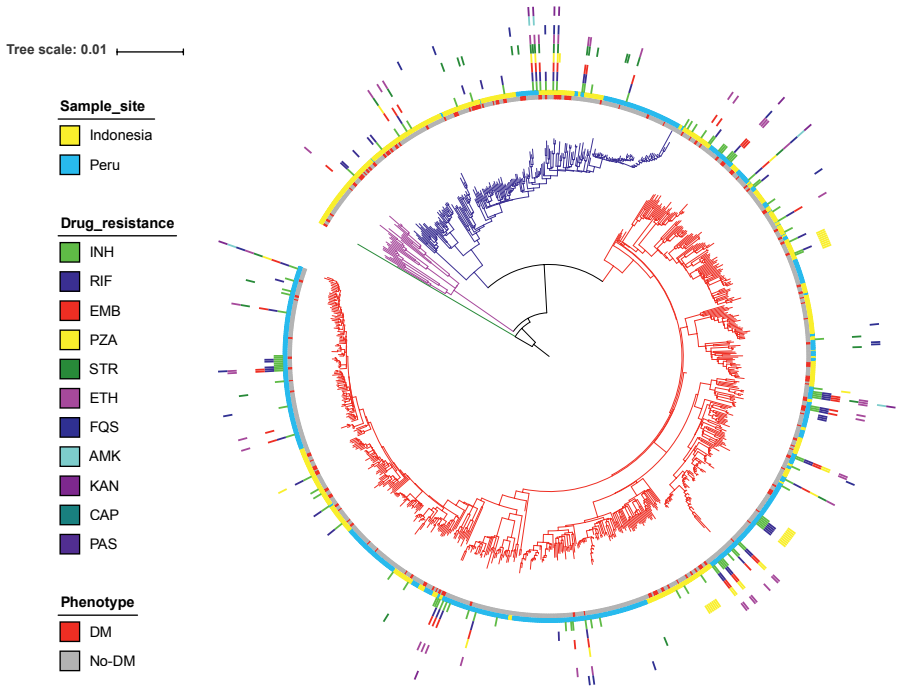


Figure 4.1. Phylogenetic tree of 896 *M. tuberculosis* isolates from TB patients with and without diabetes, showing that isolates from TB patients with diabetes were not clustered. Branch colours indicate whether isolates belong to the Indo-Oceanic (pink), East-Asian (blue), Euro-American (red) *M. tuberculosis* lineage, or to the *M. bovis* (green) lineage. The inner circle surrounding the phylogenetic tree indicates whether isolates originate from diabetes (red) or non-diabetes (grey) TB patients. The outer circle indicates isolates from Indonesian (yellow) and Peruvian (blue) patients. The bars outside the circles represent the drug resistance profile of the respective isolate.

In both countries, regardless of diabetes comorbidity, genotypic resistance was most common for isoniazid, followed by rifampicin (Table 4.3). In Peru, *Rv1482c-fabG1* (*inhA*-promoter) mutations that confer resistance to isoniazid and ethionamide, and *gyrA* mutations that confer resistance to fluoroquinolones, were significantly more common in diabetes compared to non-diabetes TB patients (Table 4.3). In Indonesia, drug resistance rates were lower and no significant differences in specific resistance mutations were observed between isolates from patients with and without diabetes, although a trend towards more frequent *rpoB* mutations was observed among TB patients with diabetes (p=0.069).

Table 4.2. Multilevel logistic regression model determining the association between diabetes and drug resistance, including a random intercept for site

	Any drug resistance (N=169)	No drug resistance (N=722)	Univariate OR (95% CI)	Adjusted OR (95% CI)
Diabetes mellitus^	38 (22%)	120 (17%)	1.6 (1.1-2.5)*	1.8 (1.1-2.9)*
Diabetes mellitus^				
No diabetes	131 (18%)	602 (82%)	REF.	REF.
Known diabetes	27 (23%)	88 (77%)	1.6 (0.96-2.6)	1.7 (0.98-2.9)
New diabetes	11 (26%)	32 (74%)	1.8 (0.9-3.6)	2.0 (0.9-4.4)
HbA1c (%) – median (IQR)^	5.7 (5.4-6.4)	5.7 (5.4-6.0)	1.0 (0.97-1.1)	1.05 (0.97-1.1)
No diabetes	5.6 (5.3-5.8)	5.6 (5.3-5.8)	1.1 (0.7-1.7)	1.4 (0.9-2.3)
Diabetes	10.5 (9.0-11.8)	11.3 (9.2-13.1)	0.9 (0.8-1.1)	0.9 (0.7-1.0)
Age <= 36 years	89 (20%)	353 (80%)	REF.	REF.
Age > 36 years	81 (18%)	372 (82%)	0.9 (0.6-1.3)	0.8 (0.5-1.1)
Male gender	92 (54%)	421 (58%)	0.8 (0.6-1.2)	0.9 (0.6-1.3)
HIV infection	2 (1%)	17 (3%)	0.5 (0.1-2.1)	0.5 (0.1-2.3)
Previous TB treatment	44 (26%)	161 (22%)	1.2 (0.8-1.8)	1.4 (0.9-2.1)
M. tuberculosis lineage				
East-Asian	34 (17%)	162 (83%)	REF.	REF.
Euro-American	135 (20%)	535 (80%)	1.2 (0.8-1.9)	1.3 (0.8-2.1)
Indo-Oceanic	1 (3%)	28 (97%)	0.2 (0.02-1.3)	0.2 (0.03-1.6)

The multivariable model was based on 816 patients without missing data (no drug resistance: n = 665; any drug resistance: n = 151). The patient with an *M. bovis* isolate was excluded from the analysis.

* P-value <0.05

^ Diabetes was included in the model as a binary (yes/no) variable, or as a categorical variable distinguishing between previously and newly diagnosed disease, or as continuous variable: HbA1c. The adjusted odds ratios represent the odds ratios from the model including the binary diabetes variable as the dependent variable and age, gender, HIV, previous TB treatment and *M. tuberculosis* lineage as independent variables.

Potential compensatory mutations

In Indonesia, three out of 37 (8%) isolates with isoniazid resistance mutations carried at least one potential isoniazid resistance-compensating mutation, compared to 39 out of 370 (11%) isolates without isoniazid resistance mutations (OR 0.7; 95% CI 0.2-2.6). In Peru, this was six out of 77 (8%) compared to nine out of 412 (2%) isolates, respectively (OR 3.8; 95% CI 1.3-11.0). Regarding rifampicin, ten out of 19 (53%) Indonesian isolates with rifampicin resistance mutations had at least one potential compensatory mutation in *rpoA*, *rpoC* and *rpoB* outside the RRDR, compared to118 out of 388 (30%) isolates without rifampicin resistance mutations

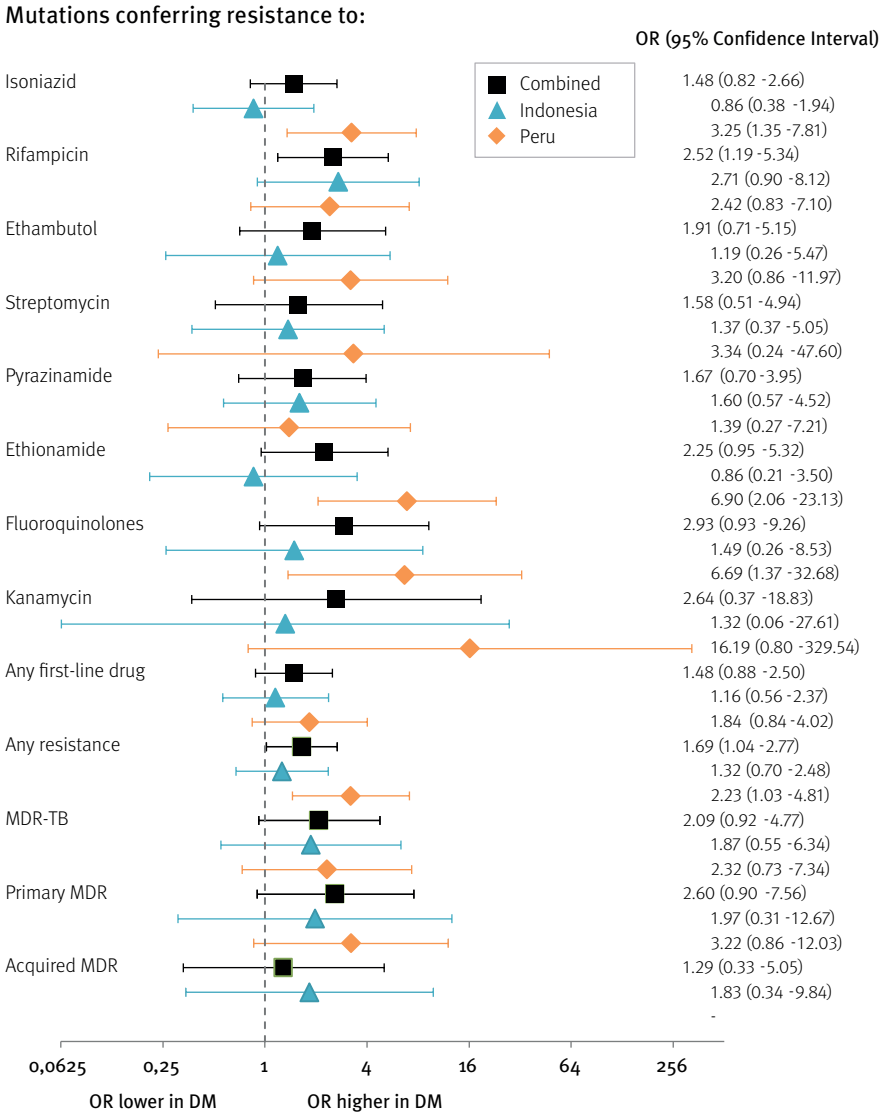


Figure 4.2. Forest plot showing the association between diabetes and genotypic drug resistance, adjusted for age, gender, site, HIV, previous treatment and *M. tuberculosis* lineage. * Odds ratios (ORs) could not be calculated for resistance to amikacin, aminoglycosides and acquired multidrug resistance (MDR) for Peru, due to too little resistant isolates. The patient with an *M. bovis* isolate was excluded from the analysis.

Table 4.3. Drug resistance genes with mutation(s) in *M. tuberculosis* isolates from patients with and without diabetes

		Indonesia		Peru	
Drug	**Gene**	**Diabetes (n=115)**	**No diabetes (n=292)**	**Diabetes (n=44)**	**No diabetes (n=445)**
Isoniazid	*katG*	8 (7.0%)	19 (6.5%)	6 (13.6%)	48 (10.8%)
	Rv1482c-fabG1	3 (2.6%)	10 (3.4%)	8 (18.2%)**	23 (5.2%)**
	ahpC	0	0	1 (2.3%)	0
Rifampicin	*rpoB*	9 (7.8%)	10 (3.4%)	7 (15.9%)	42 (9.4%)
Ethambutol	*embB*	3 (2.6%)	6 (2.1%)	5 (11.4%)	30 (6.7%)
	embC-embA	0	0	1 (2.3%)	6 (1.3%)
	embR	0	1 (0.3%)	0	0
Streptomycin	*rpsL*	4 (3.5%)	5 (1.7%)	1 (2.3%)	16 (3.6%)
	rrs	1 (0.9%)	2 (0.7%)	0	5 (1.1%)
Pyrazinamide	*pncA*	7 (6.1%)	14 (4.8%)	3 (6.8%)	22 (4.9%)
Ethionamide	*Rv1482c-fabG1*	3 (2.6%)	10 (3.4%)	8 (18.2%)**	22 (4.9%)**
	ethA	0	0	0	1 (0.2%)
Fluoroquinolones	*gyrA*	3 (2.6%)	3 (1.0%)	4 (9.1%)*	7 (1.6%)*
	gyrB	0	0	0	2 (0.4%)
Amikacin	*rrs*	0	0	0	5 (1.1%)
Capreomycin	*tlyA*	0	1 (0.3%)	0	7 (1.6%)
	rrs	0	0	0	5 (1.1%)
Kanamycin	*eis-Rv2417c*	1 (0.9%)	1 (0.3%)	1 (2.3%)	1 (0.2%)
	rrs	0	0	0	5 (1.1%)
Para-amino-salicylic acid	*folC*	0	2 (0.7%)	1 (2.3%)	2 (0.4%)
	thyA	0	0	0	1 (0.2%)

Data represent the number (%) of isolates with at least one mutation in the respective gene.
* P-value <0.05; ** P-value <0.01 (p-values are Chi-square p-values unless the expected number of resistant isolates equalled less than 5, in which case the Fisher's Exact Test p-value was calculated).

(OR 2.5; 95% CI 1.01-6.4). In Peru, this was 34 out of 49 (69%) compared to 211 out of 440 (48%) isolates, respectively (OR 2.5; 95% CI 1.3-4.6, Supplementary table S4.1). There were no differences in the association between isoniazid or rifampicin resistance mutations and potential compensatory mutations between patients with and without diabetes, or any interaction between diabetes and the presence of compensatory mutations (Supplementary table S4.1).

Sensitivity analysis

In the sensitivity analysis, distinguishing patients with newly and previously diagnosed diabetes, we confirmed the association between diabetes and drug resistance, although it was no longer significant with smaller numbers (adjusted OR 2.0; 95% CI 0.9-4.4 for those with newly, compared to OR 1.7; 95% CI 0.98-2.9 for those with previously diagnosed diabetes, Table 4.2). HbA1c was not associated with drug resistance. In addition, the association between diabetes and individual drug resistance differed per drug and per country (Figure 4.2). The association between diabetes and drug resistance remained similar after exclusion of the 19 Peruvian patients with HIV (adjusted OR 1.8; 95% CI 1.1-2.9).

Discussion

M. tuberculosis resistance mutations were more common in isolates from TB patients with diabetes, compared to those without. The association was evident for resistance to rifampicin, with a 2.5 times higher odds of rifampicin resistance in TB patients with diabetes, but was also strong for resistance to isoniazid and ethionamide, as well as fluoroquinolones, although these only reached statistical significance in the Peruvian cohort. The association of diabetes and drug resistance mutations was not explained by acquisition of drug resistance as a result of previous TB treatment since it was observed for both treatment-naïve, as well as previously treated TB patients.

Previous studies have investigated a possible link between diabetes and drug resistance, but most used phenotypic DST and focused on first-line drugs or MDR-TB (reviewed in Tegegne *et al.*[6]). Whole genome sequencing allowed us not only to investigate the association between diabetes and drug resistance at the gene level, but also to take the diverse *M. tuberculosis* genetic background into account. The association between drug resistance and diabetes has been confirmed in some[2,22], but not all[23] previous studies. One possible explanation for discrepant results could be the different populations that were studied, the size of the studies, the extent to which account was taken of factors other than diabetes status, and lack of clarity with respect to the time of diabetes testing[4]. Here, we ascertained the association between diabetes and drug resistance, adjusted for possible confounders, and discovered that the association holds true regardless of the *M. tuberculosis* strain's phylogenetic background, and was not explained by clustering of drug-resistant strains among TB patients with diabetes, as pairwise distances were not shorter among the isolates from TB patients with diabetes (Supplementary figure S4.1).

Several factors might account for the observed association between diabetes and drug resistance mutations. First, people with diabetes might be at higher risk of nosocomial transmission of drug-resistant tuberculosis in low-resource settings[24]. Second, lower rifampicin plasma concentrations among diabetes patients that were found in some[25] but not all[26] studies might lead to acquisition of drug resistance. However, all isolates in our study were collected before start of treatment, and only 26% of patients with a drug-resistant isolate reported an episode of previous TB treatment that might have resulted in acquired drug resistance mutations. Third, recent papers have found an interaction between drug resistance and cellular immunometabolism[27], and this interaction might be altered by diabetes. Fourth, similar to HIV[28], reduced host defence in people with diabetes might increase the risk of developing active TB caused by *Mycobacterium tuberculosis* strains with drug resistance mutations associated with loss of fitness[7]. Merker *et al.* have shown that the presence of mutations thought to compensate for bacterial fitness deficits was associated with transmission success and higher drug resistance rates[18]. As people with diabetes seem more likely to become infected and progress to active TB disease following infection, recently transmitted drug-resistant strains with compensatory mutations could be overrepresented among TB patients with diabetes. However, larger recent studies have challenged the paradigm of loss of fitness associated with drug resistance mutations[29,30] and we did not find differences in potential resistance-compensating mutations among TB patients with diabetes compared to those without, neither did the pairwise distances indicate more recent transmission among TB patients with diabetes. The trend towards more fluoroquinolone resistance in TB patients with diabetes could be related to the prescription of broad-spectrum antibiotics, often fluoroquinolones, in patients with respiratory symptoms who are considered at increased risk of infections due to their diabetes[31].

The current study has several limitations. First, the analysis was hindered by a 'noise' signal related to the different origin and phylogenetic backgrounds. Resistance was more common in Peru, and diabetes more common in Indonesia. However, we used multilevel regression analysis to account for this and to increase the likelihood of finding true associations. Second, we were underpowered for subgroup analyses. For instance, it is hard to draw conclusions regarding specific resistance mutations. We hope that the findings in this study may serve as a starting point for further, targeted research, powered for questions that may have arisen as a result of the presented data. Thirdly, we could not prove if diabetes is associated with more transmission of drug-resistance as the sampling fraction of *M. tuberculosis* isolates in the two high-endemic settings was likely to be too low, making it difficult to identify transmission clusters. Besides higher rates of transmission, diabetes may also lead to more TB reactivation caused by drug-resistant strains.

In summary, for the first time, we used *M. tuberculosis* whole genome sequencing data from two large cohorts of TB patients with and without diabetes to study the association between diabetes and genotypic drug resistance in TB patients. Diabetes was associated with an increased risk of disease caused by strains with resistance mutations, particularly those against rifampicin, but also against isoniazid, ethionamide and fluoroquinolones. Higher rates of resistance mutations among TB patients with diabetes were not explained by acquisition of drug resistance mutations during treatment, but might be related to increased susceptibility to opportunistic infection with drug-resistant TB. TB patients with diabetes should be prioritized for DST in settings where this is not performed for all patients.

Acknowledgements

We thank the Oxford Genomics Centre at the Welcome Centre for Human Genetics (funded by Welcome Trust grant reference 203141/Z/16/Z) for the generation and initial processing of the whole genome sequencing data.

Funding

This work was supported by the TANDEM project, which was funded by the European Union's Seventh Framework Programme (FP7/2007–2013) under Grant Agreement Number 305279.

Whole genome sequencing: CRyPTIC consortium, funded by a Welcome Trust/Newton Fund-MRC Collaborative Award (200205/Z/15/Z) and the Bill and Melinda Gates Foundation Trust (OPP1133541).

Supplementary methods

Whole genome sequencing

M. tuberculosis DNA was extracted after subculturing positive cultures on Ogawa (Indonesia) or Middlebrook 7H10-7H11 (Peru) solid medium using the cetyl trimethylammonium bromide (CTAB) method or the UltraClean® Microbial DNA Isolation Kit (MO BIO Laboratories, Carlsbad, CA)[32].

M. tuberculosis DNA from 148 Indonesian isolates was sequenced on an Illumina HiSeq 2000 instrument using 2 x 100 bp paired-end reads at the Beijing Genome Institute in Hong Kong, and on an Illumina NextSeq500 instrument using 2 x 150 bp paired-end reads at the department of Human Genetics at the Radboudumc, Nijmegen, the Netherlands for the remaining 259 Indonesian isolates. *M. tuberculosis* DNA from Peruvian isolates was sequenced on an Illumina HiSeq 4000 instrument using 2 x 150 bp paired-end reads at the Wellcome Centre for Human Genetics in Oxford, United Kingdom. Sequencing coverage was determined using the FastQC quality control tool version 0.10.1, and the Genome Analysis Toolkit[33]. The average coverage depth for the 896 isolates was 110.7x, and the average percentage of bases covered by at least one read was 98.7%. Nine isolates failed the sequencing or mapping quality check and were excluded from further analysis. Sequencing quality control statistics are shown in Supplementary file 1.

We excluded the *rrs* C492T mutation from the resistance mutation database, because we considered this mutation to be a phylogenetic marker of the LAM3 sublineage rather than a resistance mutation[34]. We used an allele frequency cut-off of ≥ 30%, as previously proposed[35]. In addition, to identify potential compensatory mutations, variants were called with Breseq software, version 0.27.1[36] with a minimum threshold of 30x coverage. Mutations with low-quality evidence (i.e. possible mixed read alignment) were not included. We reported mutations according to the *M. tuberculosis* complex numbering system, based on the sequence of the reference strain, *M. tuberculosis* H37Rv[37].

Phylogeny construction

For phylogeny construction, sequence reads were aligned to reference strain *M. tuberculosis* H37Rv, accession number NC_000962.3, and variants were called using Breseq, version 0.27.1[36]. We extracted all 58,041 variable positions across the 896 *M. tuberculosis* sequences and concatenated them into a multiple sequence alignment. Solely for the purpose of creating the phylogenetic tree, SNPs occurring in PE/PPE genes, genes related to mobile elements, as well as genes previously associated with drug resistance[38] were removed. The remaining 56,654 SNPs were

used to construct the phylogenetic tree using PhyML version 3.0[19] using the HKY85 model with four gamma-distributed rate-categories, and using a hundred bootstraps.

Statistical analysis

We used the glmer() function from the lme4 library[39] to assess the association between diabetes and genotypic drug resistance. We used a multilevel regression model with robust standard errors to account for clustering within countries[40]. Diabetes, age, gender, HIV-infection, previous TB treatment, and *M. tuberculosis* lineage were included as categorical independent variables in the model. Age was included as a binary variable with the median age (36 years) as cut-off value for the best fit. The single patient with an *M. bovis* isolate was excluded from the multilevel logistic regression analysis, because it added complexity to the model and did not improve model fitness. Data were analysed using complete case analysis.

Significance of possible differences in resistance mutations between diabetes and non-diabetes TB patients were determined with the Fisher's exact test or Chi-square test if there were more than two categories per variable. To assess the association between isoniazid and rifampicin resistance mutations and the presence of previously proposed accompanying compensatory mutations, stratified for diabetes comorbidity, for Indonesian and Peruvian isolates separately, we compared crude odds ratios with stratum-specific and Mantel-Haenszel odds ratios. The Breslow-Day test was used to test the significance of interaction between diabetes and the presence of potential compensatory mutations. All analyses were performed in R, version 3.5.2.

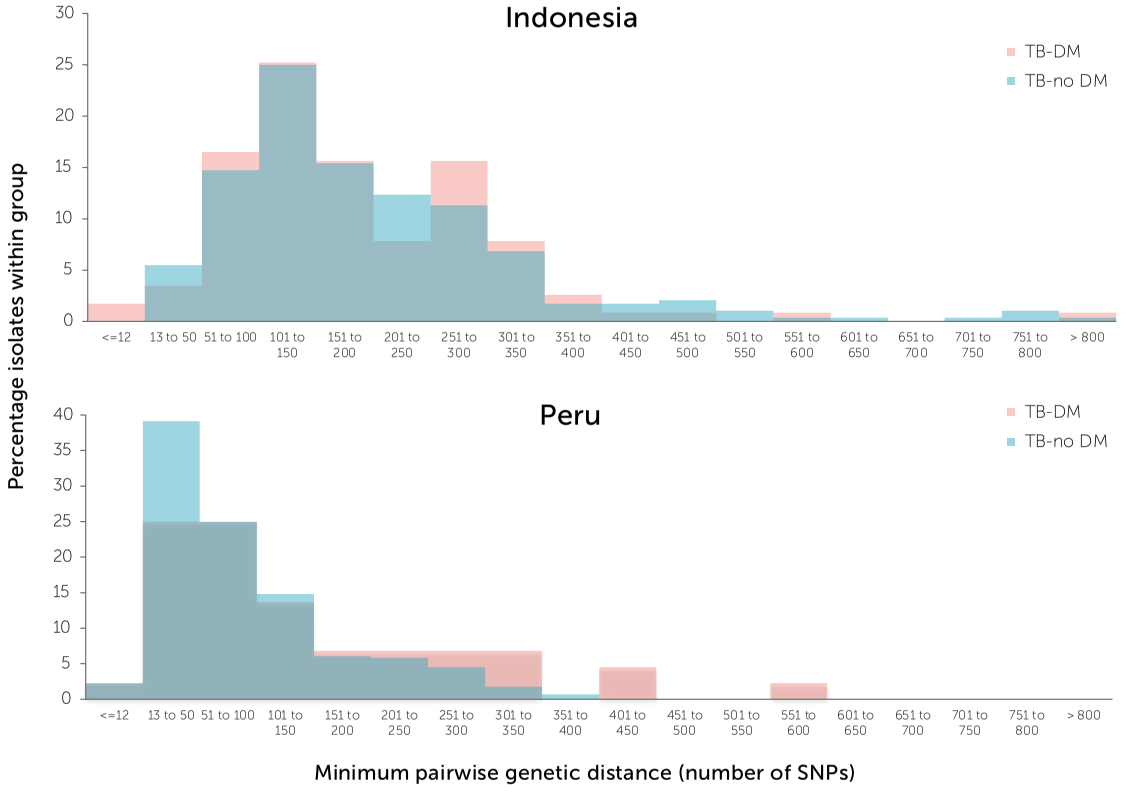


Figure S4.1. Histograms of minimum pairwise genetic distances for *M. tuberculosis* isolates from patients with and without diabetes.

Table S4.1. Association between putative compensatory mutations and isoniazid and rifampicin resistance, stratified by diabetes comorbidity

			INH-resistant	INH- susceptible	Odds ratio (95% CI)	Crude & Mantel-Haenszel OR
Indonesia	No diabetes	Compensatory mutation(s)	3	31	0.9 (0.3-3.3)	0.7 (0.2-2.6)
		No compensatory mutation	24	234		&
	Diabetes	Compensatory mutation(s)	0	8	0.6 (0.03-10.2)	0.7 (0.2-2.5)
		No compensatory mutation	10	97		
Peru	No diabetes	Compensatory mutation(s)	5	8	4.0 (1.3-12.5)*	3.8 (1.3-11.0)*
		No compensatory mutation	59	373		&
	Diabetes	Compensatory mutation(s)	1	1	2.5 (0.1-43.4)	3.7 (1.2-10.7)*
		No compensatory mutation	12	30		
			RIF-resistant	RIF-susceptible	Odds ratio (95% CI)	Crude & Mantel-Haenszel OR
Indonesia	No diabetes	Compensatory mutation(s)	5	91	2.1 (0.6-7.4)	2.5 (1.01-6.4)*
		No compensatory mutation	5	191		&
	Diabetes	Compensatory mutation(s)	5	27	3.7 (0.9-14.6)	2.7 (1.1-6.8)*
		No compensatory mutation	4	79		
Peru	No diabetes	Compensatory mutation(s)	29	195	2.4 (1.2-4.7)*	2.5 (1.3-4.6)*
		No compensatory mutation	13	208		&
	Diabetes	Compensatory mutation(s)	5	16	3.3 (0.6-19.2)	2.5 (1.3-4.7)*
		No compensatory mutation	2	21		
Sites combined	No diabetes	Compensatory mutation(s)	34	286	2.6 (1.5-4.8)*	2.8 (1.7-4.7)*
		No compensatory mutation	18	399		&
	Diabetes	Compensatory mutation(s)	10	43	3.9 (1.3-11.3)*	2.9 (1.7-4.8)*
		No compensatory mutation	6	100		

NOTE. Isoniazid and rifampicin resistance were defined as any mutations in genes known to confer resistance to isoniazid and rifampicin according to TB Profiler (19). For the top-half of the table, relating to isoniazid resistance, compensatory mutations concerned any mutations in *ubiA* and the *ahpC* promotor region; for the bottom-half of the table, relating to rifampicin resistance, compensatory mutations concerned any mutations in *rpoA*, *rpoB* outside the rifampicin resistance-determining region, and *rpoC*. * P-value ‹0.05

References

1. Pan S-C, Ku C-C, Kao D, Ezzati M, Fang C-T, Lin H-H. Effect of diabetes on tuberculosis control in 13 countries with high tuberculosis: a modelling study. The Lancet Diabetes & Endocrinology 2015: 3(5): 323-330.
2. Perez-Navarro LM, Restrepo BI, Fuentes-Dominguez FJ, Duggirala R, Morales-Romero J, Lopez-Alvarenga JC, Comas I, Zenteno-Cuevas R. The effect size of type 2 diabetes mellitus on tuberculosis drug resistance and adverse treatment outcomes. Tuberculosis (Edinb) 2017: 103: 83-91.
3. Liu Q, Li W, Xue M, Chen Y, Du X, Wang C, Han L, Tang Y, Feng Y, Tao C, He JQ. Diabetes mellitus and the risk of multidrug resistant tuberculosis: a meta-analysis. Sci Rep 2017: 7(1): 1090.
4. Huangfu PU-G, C.; Golub, J.; Pearson, F.; Critchley, J. The effects of diabetes on tuberculosis treatment outcomes: an updated systematic review and meta-analysis. INTERNATIONAL JOURNAL OF TUBERCULOSIS AND LUNG DISEASE 2019.
5. Riza AL, Pearson F, Ugarte-Gil C, Alisjahbana B, van de Vijver S, Panduru NM, Hill PC, Ruslami R, Moore D, Aarnoutse R, Critchley JA, van Crevel R. Clinical management of concurrent diabetes and tuberculosis and the implications for patient services. The Lancet Diabetes & Endocrinology 2014: 2(9): 740-753.
6. Tegegne BS, Mengesha MM, Teferra AA, Awoke MA, Habtewold TD. Association between diabetes mellitus and multi-drug-resistant tuberculosis: evidence from a systematic review and meta-analysis. Syst Rev 2018: 7(1): 161.
7. Dheda K, Gumbo T, Maartens G, Dooley KE, McNerney R, Murray M, Furin J, Nardell EA, London L, Lessem E, Theron G, van Helden P, Niemann S, Merker M, Dowdy D, Van Rie A, Siu GKH, Pasipanodya JG, Rodrigues C, Clark TG, Sirgel FA, Esmail A, Lin H-H, Atre SR, Schaaf HS, Chang KC, Lange C, Nahid P, Udwadia ZF, Horsburgh CR, Churchyard GJ, Menzies D, Hesseling AC, Nuermberger E, McIlleron H, Fennelly KP, Goemaere E, Jaramillo E, Low M, Jara CM, Padayatchi N, Warren RM. The epidemiology, pathogenesis, transmission, diagnosis, and management of multidrug-resistant, extensively drug-resistant, and incurable tuberculosis. The Lancet Respiratory Medicine 2017: 5(4): 291-360.
8. Walker TM, Merker M, Kohl TA, Crook DW, Niemann S, Peto TE. Whole genome sequencing for M/XDR tuberculosis surveillance and for resistance testing. Clin Microbiol Infect 2017: 23(3): 161-166.
9. van Crevel R, Dockrell HM. TANDEM: understanding diabetes and tuberculosis. The Lancet Diabetes & Endocrinology 2014: 2(4): 270-272.
10. Grint D, Alisjhabana B, Ugarte-Gil C, Riza AL, Walzl G, Pearson F, Ruslami R, Moore DAJ, Ioana M, McAllister S, Ronacher K, Koeseomadinata RC, Kerry-Barnard SR, Coronel J, Malherbe ST, Dockrell HM, Hill PC, Van Crevel R, Critchley JA, consortium T. Accuracy of diabetes screening methods used for people with tuberculosis, Indonesia, Peru, Romania, South Africa. Bull World Health Organ 2018: 96(11): 738-749.
11. Ugarte-Gil C, Alisjahbana B, Ronacher K, Riza AL, Koesoemadinata RC, Malherbe ST, Cioboata R, Llontop JC, Kleynhans L, Lopez S, Santoso P, Marius C, Villaizan K, Ruslami R, Walzl G, Panduru NM, Dockrell HM, Hill PC, Mc Allister S, Pearson F, Moore DAJ, Critchley JA, van Crevel R, Consortium T. Diabetes mellitus among pulmonary tuberculosis patients from four TB-endemic countries: the TANDEM study. Clin Infect Dis 2019.
12. WHO. Use of glycated haemoglobin (HbA1c) in diagnosis of diabetes mellitus: abbreviated report of a WHO consultation. Geneva, Switzerland: World Health Organization; 2011.
13. Huangfu P, Laurence YV, Alisjahbana B, Ugarte-Gil C, Riza AL, Walzl G, Ruslami R, Moore DAJ, Ioana M, McAllister S, Ronacher K, Koesoemadinata RC, Grint D, Kerry S, Coronel J, Malherbe ST, Griffiths U, Dockrell HM, Hill PC, van Crevel R, Pearson F, Critchley JA. Point of care HbA1c level for diabetes mellitus management and its accuracy among tuberculosis patients: a study in four countries. Int J Tuberc Lung Dis 2019: 23(3): 283-292.
14. Critchley JA. Appendices to "Diabetes screening in tuberculosis patients: a diagnostic accuracy analysis of risk scores and laboratory methods in Indonesia, Peru, Romania and South Africa". 2018.
15. Coll F, Preston M, Guerra-Assuncao JA, Hill-Cawthorn G, Harris D, Perdigao J, Viveiros M, Portugal I, Drobniewski F, Gagneux S, Glynn JR, Pain A, Parkhill J, McNerney R, Martin N, Clark TG. PolyTB: a genomic variation map for Mycobacterium tuberculosis. Tuberculosis (Edinb) 2014: 94(3): 346-354.

16. Sherman DR, Mdluli K, Hickey MJ, Arain TM, Morris SL, Barry CE, 3rd, Stover CK. Compensatory ahpC gene expression in isoniazid-resistant Mycobacterium tuberculosis. Science 1996: 272(5268): 1641-1643.
17. Cohen KA, Abeel T, Manson McGuire A, Desjardins CA, Munsamy V, Shea TP, Walker BJ, Bantubani N, Almeida DV, Alvarado L, Chapman SB, Mvelase NR, Duffy EY, Fitzgerald MG, Govender P, Gujja S, Hamilton S, Howarth C, Larimer JD, Maharaj K, Pearson MD, Priest ME, Zeng Q, Padayatchi N, Grosset J, Young SK, Wortman J, Mlisana KP, O'Donnell MR, Birren BW, Bishai WR, Pym AS, Earl AM. Evolution of Extensively Drug-Resistant Tuberculosis over Four Decades: Whole Genome Sequencing and Dating Analysis of Mycobacterium tuberculosis Isolates from KwaZulu-Natal. PLoS Med 2015: 12(9): e1001880.
18. Merker M, Barbier M, Cox H, Rasigade JP, Feuerriegel S, Kohl TA, Diel R, Borrell S, Gagneux S, Nikolayevskyy V, Andres S, Nubel U, Supply P, Wirth T, Niemann S. Compensatory evolution drives multidrug-resistant tuberculosis in Central Asia. Elife 2018: 7.
19. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 2010: 59(3): 307-321.
20. Ronacher K, van Crevel R, Critchley JA, Bremer AA, Schlesinger LS, Kapur A, Basaraba R, Kornfeld H, Restrepo BI. Defining a Research Agenda to Address the Converging Epidemics of Tuberculosis and Diabetes. Chest 2017: 152(1): 174-180.
21. Naidoo CC, Pillay M. Fitness-compensatory mutations facilitate the spread of drug-resistant F15/LAM4/KZN and F28 Mycobacterium tuberculosis strains in KwaZulu-Natal, South Africa. Journal of Genetics 2017: 96(4): 599-612.
22. Hsu AH, Lee JJ, Chiang CY, Li YH, Chen LK, Lin CB. Diabetes is associated with drug-resistant tuberculosis in Eastern Taiwan. Int J Tuberc Lung Dis 2013: 17(3): 354-356.
23. Leung CC, Yew WW, Mok TYW, Lau KS, Wong CF, Chau CH, Chan CK, Chang KC, Tam G, Tam CM. Effects of diabetes mellitus on the clinical presentation and treatment response in tuberculosis. Respirology 2017: 22(6): 1225-1232.
24. van Crevel R, van de Vijver S, Moore DAJ. The global diabetes epidemic: what does it mean for infectious diseases in tropical countries? The Lancet Diabetes & Endocrinology 2017: 5(6): 457-468.
25. Alkabab Y, Keller S, Dodge D, Houpt E, Staley D, Heysell S. Early interventions for diabetes related tuberculosis associate with hastened sputum microbiological clearance in Virginia, USA. BMC Infect Dis 2017: 17(1): 125.
26. Ruslami R, Nijland HM, Adhiarta IG, Kariadi SH, Alisjahbana B, Aarnoutse RE, van Crevel R. Pharmacokinetics of antituberculosis drugs in pulmonary tuberculosis patients with type 2 diabetes. Antimicrob Agents Chemother 2010: 54(3): 1068-1074.
27. Howard NC, Marin ND, Ahmed M, Rosa BA, Martin J, Bambouskova M, Sergushichev A, Loginicheva E, Kurepina N, Rangel-Moreno J, Chen L, Kreiswirth BN, Klein RS, Balada-Llasat JM, Torrelles JB, Amarasinghe GK, Mitreva M, Artyomov MN, Hsu FF, Mathema B, Khader SA. Mycobacterium tuberculosis carrying a rifampicin drug resistance mutation reprograms macrophage metabolism through cell wall lipid changes. Nat Microbiol 2018: 3(10): 1099-1108.
28. Mesfin YM, Hailemariam D, Biadgilign S, Kibret KT. Association between HIV/AIDS and multi-drug resistance tuberculosis: a systematic review and meta-analysis. PLoS One 2014: 9(1): e82235.
29. Becerra MC, Huang C-C, Lecca L, Bayona J, Contreras C, Calderon R, Yataco R, Galea J, Zhang Z, Atwook S, Cohen T, Mitnick CD, Farmer P, Murray M. 2018.
30. Liu Q, Zuo T, Xu P, Jiang Q, Wu J, Gan M, Yang C, Prakash R, Zhu G, Takiff HE, Gao Q. Have compensatory mutations facilitated the current epidemic of multidrug-resistant tuberculosis? Emerg Microbes Infect 2018: 7(1): 98.
31. Long R, Chong H, Hoeppner V, Shanmuganathan H, Kowalewska-Grochowska K, Shandro C, Manfreda J, Senthilselvan A, Elzainy A, Marrie T. Empirical treatment of community-acquired pneumonia and the development of fluoroquinolone-resistant tuberculosis. Clin Infect Dis 2009: 48(10): 1354-1360.
32. Somerville W, Thibert L, Schwartzman K, Behr MA. Extraction of Mycobacterium tuberculosis DNA: a question of containment. J Clin Microbiol 2005: 43(6): 2996-2997.
33. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010: 20(9): 1297-1303.

34. Tudo G, Rey E, Borrell S, Alcaide F, Codina G, Coll P, Martin-Casabona N, Montemayor M, Moure R, Orcau A, Salvado M, Vicente E, Gonzalez-Martin J. Characterization of mutations in streptomycin-resistant Mycobacterium tuberculosis clinical isolates in the area of Barcelona. J Antimicrob Chemother 2010: 65(11): 2341-2346.
35. Black PA, de Vos M, Louw GE, van der Merwe RG, Dippenaar A, Streicher EM, Abdallah AM, Sampson SL, Victor TC, Dolby T, Simpson JA, van Helden PD, Warren RM, Pain A. Whole genome sequencing reveals genomic heterogeneity and antibiotic purification in Mycobacterium tuberculosis isolates. BMC Genomics 2015: 16(1): 857.
36. Barrick JE. Identifying structural variation in haploid microbial genomes from short-read resequencing data using breseq. BMC Genomics 2014.
37. Andre E, Goeminne L, Cabibbe A, Beckert P, Kabamba Mukadi B, Mathys V, Gagneux S, Niemann S, Van Ingen J, Cambau E. Consensus numbering system for the rifampicin resistance-associated rpoB gene mutations in pathogenic mycobacteria. Clin Microbiol Infect 2017: 23(3): 167-172.
38. Coll F, McNerney R, Preston MD, Guerra-Assuncao JA, Warry A, Hill-Cawthorne G, Mallard K, Nair M, Miranda A, Alves A, Perdigao J, Viveiros M, Portugal I, Hasan Z, Hasan R, Glynn JR, Martin N, Pain A, Clark TG. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Med 2015: 7(1): 51.
39. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Usinglme4. Journal of Statistical Software 2015: 67(1).
40. Austin PC, Merlo J. Intermediate and advanced topics in multilevel logistic regression analysis. Stat Med 2017: 36(20): 3257-3277.

PART TWO

Tuberculosis disease phenotype and transmission

5

Large-scale genomic analysis shows association between homoplastic genetic variation in *Mycobacterium tuberculosis* genes and meningeal or pulmonary tuberculosis

Carolien Ruesen, Lidya Chaidir, Arjan van Laarhoven, Sofiati Dian, Ahmad Rizal Ganiem, Hanna Nebenzahl-Guimaraes, Martijn A. Huynen, Bachti Alisjahbana, Bas E. Dutilh, Reinout van Crevel.

BMC Genomics. 2018; 19:122.

Abstract

Background

Meningitis is the most severe manifestation of tuberculosis. It is largely unknown why some people develop pulmonary TB (PTB) and others TB meningitis (TBM); we examined if the genetic background of infecting *M. tuberculosis* strains may be relevant.

Methods

We whole-genome sequenced *M. tuberculosis* strains isolated from 322 HIV-negative tuberculosis patients from Indonesia and compared isolates from patients with TBM (n = 106) and PTB (n = 216). Using a phylogeny-adjusted genome-wide association method to count homoplasy events we examined phenotype-related changes at specific loci or genes in parallel branches of the phylogenetic tree. Enrichment scores for the TB phenotype were calculated on single nucleotide polymorphism (SNP), gene, and pathway level. Genetic associations were validated in an independent set of isolates.

Results

Strains belonged to the East-Asian lineage (36.0%), Euro-American lineage (61.5%), and Indo-Oceanic lineage (2.5%). We found no association between lineage and phenotype (Chi-square = 4.556; p = 0.207). Large genomic differences were observed between isolates; the minimum pairwise genetic distance varied from 17 to 689 SNPs. Using the phylogenetic tree, based on 28,544 common variable positions, we selected 54 TBM and 54 PTB isolates in terminal branch sets with distinct phenotypes. Genetic variation in *Rv0218*, and absence of *Rv3343c*, and *nanK* were significantly associated with disease phenotype in these terminal branch sets, and confirmed in the validation set of 214 unpaired isolates.

Conclusions

Using homoplasy counting we identified genetic variation in three separate genes to be associated with the TB phenotype, including one (*Rv0218*) which encodes a secreted protein that could play a role in host-pathogen interaction by altering pathogen recognition or acting as virulence effector.

Introduction

Tuberculosis (TB), caused by *Mycobacterium tuberculosis*, remains a major global health problem[1]. Active TB mostly affects the lungs but may also spread to other organs. TB meningitis (TBM), which represents approximately 1–5% of all TB cases, is the most severe manifestation of TB, resulting in death or neurological disability in about half of those affected[2,3]. It is largely unknown why certain people develop pulmonary TB (PTB) and others TBM. Host immune-related factors clearly play an important role, as shown by the increased risk of TBM for patients with advanced HIV infection, and the overrepresentation of young children among TBM patients. Host genetic factors may also play a role; single studies have linked susceptibility to TBM with variation in candidate genes[4–8].

Besides the host, genetic diversity of infecting *M. tuberculosis* strains may also affect disease phenotype. Even though *M. tuberculosis* is considered a clonal organism, there is considerable genetic variation in the genomes of infecting *M. tuberculosis* isolates[9,10]. Epidemiological studies have reported significant differences among *M. tuberculosis* lineages in terms of virulence[11,12], transmission[9,13,14], progression to active disease after infection[15], and response to treatment[16,17]. *In vitro* studies have supported these findings by showing *M. tuberculosis* genotype-specific differences in the human immune response[18–21].

Animal studies have shown that *M. tuberculosis* strains differ in their ability to invade the central nervous system (CNS). Five *M. tuberculosis* genes (*Rv0311* (unknown function), *Rv0805* (intermediary metabolism and respiration), *pknD* (protein kinase D), *Rv0986* (cell wall and cell processes), and MT3280 (unknown function)) have been associated with invasion or survival in the CNS but not in lung tissues in mice[22]. Especially *M. tuberculosis pknD* was associated with invasion of brain, but not lung epithelia in guinea pigs[23], as was confirmed by another study showing that *pknD* vaccination offered significant protection against bacterial dissemination to the brain in guinea pigs[24]. Similarly, in mice, clinical isolates from TBM patients disseminated extensively to cause meningitis, whereas *M. tuberculosis* H37Rv and clinical isolates from PTB patients did not[25]. In rabbits, production of phenolic glycolipid has been linked with the increased propensity of East-Asian/Beijing strains to cause TBM[26]. Finally, four *M. tuberculosis* genes were crucial for invading an artificial blood brain barrier in an *in vitro* model using primary human brain microvascular endothelial cells: *PE-PGRS18* (unknown function), *Rv0987* (cell wall and cell processes), *grcC2* (intermediary metabolism and respiration), and *PPE29* (unknown function)[27].

Much less is known about the role of *M. tuberculosis* genotype in TBM in humans. Most studies have examined associations of *M. tuberculosis* lineage with disease phenotype. Compared to other lineages, strains belonging to the East-Asian lineage were associated with extrapulmonary tuberculosis in one study[28], but not in another[29], while other studies found no association of *M. tuberculosis* lineage and disease localisation[30,31]. Specifically looking at TBM, one study from Vietnam found the Euro-American lineage to be associated with PTB rather than TBM[32]. Only one study used whole genome sequencing to compare strains from TBM and PTB patients; large-scale and smaller genomic rearrangements, inversions, indels and single nucleotide polymorphisms (SNPs) in eight cerebrospinal fluid (CSF)-derived strains were not found in 69 comparison respiratory strains isolated from independent sputum samples[33]. In the current study, we used a much larger set of isolates and a novel approach to examine the effect of the *M. tuberculosis* genotype on the susceptibility to TBM. We compared *M. tuberculosis* genomes isolated from 216 PTB patients and 106 TBM patients from Indonesia, all HIV-negative, to detect homoplastic genetic variants associated with either PTB or TBM.

Results

Lineage distribution and phylogeny construction

M. tuberculosis isolates from established patient cohorts in Bandung, Indonesia were selected for whole genome sequencing. All available *M. tuberculosis* strains isolated from HIV-negative TBM patients and randomly selected strains from twice as many PTB patients from the same setting were included, one strain was selected per patient. Compared to the 216 PTB patients, the 106 TBM patients were from a similar ethnic background, but slightly younger, more often male, and more often previously treated for TB (Table S5.1). Based on a 62-SNP barcode[34] 61.5% of the strains belonged to the Euro-American lineage (63.4% for PTB; 57.5% for TBM), 36% to the East-Asian lineage (33.3% for PTB; 41.5% for TBM), and 2.5% to the Indo-Oceanic lineage (3.2% for PTB; 0.9% for TBM). The lineage distribution did not differ significantly for strains isolated from TBM compared to PTB patients (Chi-square = 3.230; p = 0.199).

A phylogenetic tree was constructed based on 28,544 variable common nucleotide positions among the 322 *M. tuberculosis* isolates. The phylogeny showed that the TBM phenotype was not restricted to a certain *M. tuberculosis* lineage; instead it arose many times independently (Figure 5.1). In addition, the tree showed a high degree of strain heterogeneity. On average, two strains differed by about 1,000 SNPs, and this pairwise distance did not differ among PTB and TBM strains (data not shown), indicating that there was equal genetic diversity within PTB and within TBM strains. In

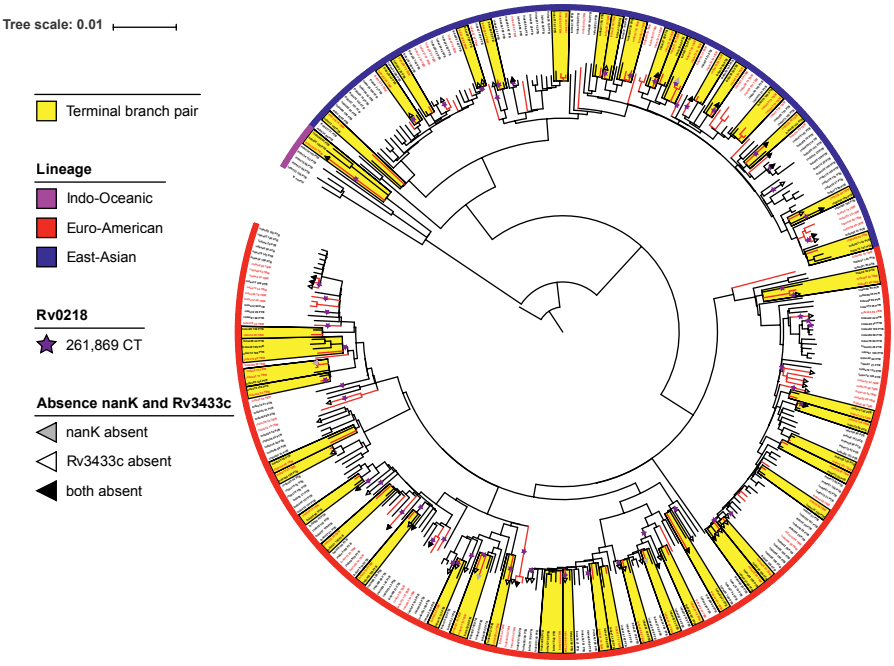


Figure 5.1. Phylogenetic tree of 322 *M. tuberculosis* strains isolated from TBM (red braches and leaves) and PTB (black branches and leaves) patients. The highlighted branches indicate the 108 strains in 47 terminal branch sets, together comprising the discovery set. The purple stars indicate the origin of the SNP in *Rv0218* according to the ancestral reconstruction. The grey and black triangles indicate the isolates in which *nanK* and/or *Rv3433c* are absent.

addition, the minimum pairwise genetic distance varied from 17 to 689 SNPs (data not shown), indicating that there was no clustering of strains (≤12 SNPs distance[35]). For TBM strains the minimum genetic distance ranged from 17 to 1,785, and for PTB strains from 31 to 803 SNPs (data not shown).

TB phenotype-associated genetic variations

Genome-wide association approaches for bacteria can broadly be categorised into allele counting and homoplasy counting methods[36]. Allele counting methods are based on the overrepresentation of an allele at the same site in cases relative to controls, introducing a risk of false-positive findings due to population stratification. Homoplasy counting on the other hand, counts repeated and independently emerging mutations that occur more often in branches of cases relative to controls. In the current study, we used a two-step approach: in the discovery phase, we used homoplasy

counting by identifying terminal branch sets (TBSs) to maximize power to identify true associations with the TB disease phenotype, uncorrected for multiple tests. In the validation phase we examined associations identified in the discovery phase using allele counting with correction for multiple testing and for phylogenetic bias to distinguish true associations from false positives, and performed ancestral reconstruction to remove possible phylogenetic bias. To divide the genomes in a discovery and a validation set, we identified isolates in terminal branch pairs, trios and quartets (i.e. separated at a terminal or near terminal branch in the phylogenetic tree) with distinct phenotypes (Figure 5.2). Genetic differences between isolates within a TBS provide the strongest, homoplasy-corrected possible association with the phenotype.

The phylogenetic tree revealed a total of 47 TBSs containing 108 paired strains: 54 TBM and 54 PTB strains that make up the discovery set. The merged SNP lists consisted of 6,488 variable positions with the corresponding nucleotides for the 108 strains (Additional file 2). Using the homoplasy-based association analysis, we found individual nucleotide positions, genes, or pathways where TB disease phenotype-associated mutations repeatedly occur in different branches of the phylogenetic tree. These included 9 SNPs, 5 genes, and 1 pathway (Table 5.1).

We used the remaining 214 (52 TBM and 162 PTB) isolates not belonging to any of the TBSs (validation set) to verify these results. The discovery set showed a total of 6,488 different non-synonymous SNPs involving 6,483 dimorphic sites and 5 trimorphic sites across 2,778 genes; the validation set a total of 12,211 different nonsynonymous SNPs involving 12,185 dimorphic sites and 26 trimorphic sites across 3,359 genes (Additional file 3). There was an overlap of 2,694 non-synonymous SNPs and 1,564 affected genes. Out of 9 SNPs significantly associated with either TBM or PTB in the discovery set, one was confirmed in the validation set; the mutation in *Rv0218*. Similarly, out of 5 genes harbouring genetic variation associated with the TB phenotype, *Rv0218* was validated in the validation set (Table 5.1, Figure S5.1 and Figure S5.2). The pathway (ethylbenzene degradation) identified in the discovery set was not confirmed in the validation set.

To correct for potential phylogenetic bias in the validation set, we reconstructed the ancestral state for the SNP in *Rv0218* and compared the ratio of TBM vs. PTB isolates after the occurrence of this particular SNP with the ratio of TBM vs. PTB prior to the occurrence of this SNP in the validation set (Figure 5.1 and Figure S5.3). Three branches with back mutations (2 TBM, 1 PTB branch) were excluded from the analysis. Among the 33 nodes / leaves where the SNP occurred, the average unweighted proportion of TBM isolates among the child branches was 44.7%; among 166 isolates in the validation set not harbouring this SNP, 29 (17.5%) were from TBM patients (Table S5.2). The Z-score for the difference between proportions was –3.83 ($p < 0.001$).

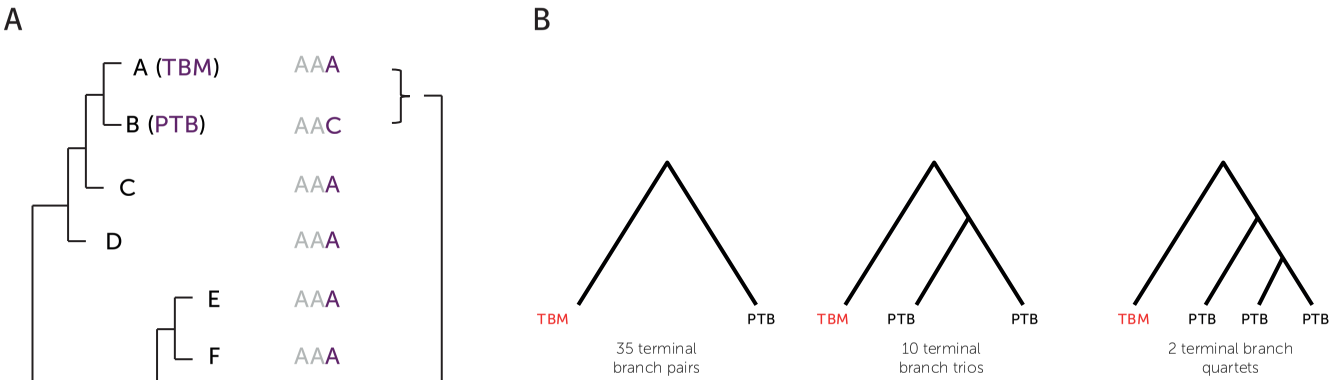


Figure 5.2. Homoplasy-based association analysis; detection of mutations occurring along disparate locations in the phylogenetic tree. Left part (A): schematic example of a tree containing two terminal branch sets (TBSs) with a homoplastic SNP. Right part (B): schematic overview of the composition of the 47 terminal branch sets.

Table 5.1. Significant SNPs, genes, and pathways identified by homoplasy counting.

		Discovery dataset (N=108)			Validation dataset (N=214)		
SNP-level		Strains with SNP (N)			Strains with SNP (N)		
Gene (Rv-number)	Nucleotide change	TBM (N=54)	PTB (N=54)	P-value	TBM (N=52)	PTB (N=162)	P-value^
Unnamed (*Rv0218*)	261,869CT	25	10	0.002	21	24	0.001
PPE54 (*Rv3343c*)	3,736,628TG	22	35	0.008	27	87	0.472
PEPGRS19 (*Rv1067c*)	1,190,093AC	2	11	0.01	5	18	0.487
PEPGRS44 (*Rv2591*)	2,922,848AT	0	6	0.01	2	7	0.625
PPE3 (*Rv0280*)	340,372TC	2	10	0.025	4	13	0.589
PEPGRS9 (*Rv0746*)	836,272AG	7	1	0.029	3	9	0.603
PEPGRS26 (*Rv1441c*)	16,18,978TC	0	5	0.032	2	3	0.369
Unnamed (*Rv0064*)	713,36GC	24	34	0.034	29	98	0.355
PEPGRS18 (*Rv0980c*)	1,095,644CT	0	4	0.044	2	2	0.238
Gene-level		SNPs in gene (N)			SNPs in gene (N)		
Gene (Rv number)		TBM	PTB	P-value	TBM	PTB	P-value
PEPGRS19 (*Rv1067c*)		2	13	0.004	5	22	0.318
Unnamed (*Rv0218*)		27	11	0.007	21	32	0.001
PEPGRS26 (*Rv1441c*)		0	5	0.031	2	4	0.491
glmS (*Rv3436c*)		0	5	0.031	0	0	n.a.
Unnamed (*Rv3740c*)		0	4	0.05	3	0	0.011
Pathway-level		Genes with mutation (N)			Genes with mutation (N)		
Pathway name		TBM	PTB	P-value	TBM	PTB	P-value
Ethylbenzene degradation		68	54	0.032	65	224	0.071

NOTE. P-values are based on permutation analysis; bold p-values indicate validated, Bonferroni-corrected significant enrichment ^ P-value thresholds in the validation set were Bonferroni-corrected for multiple testing by dividing them by the number of top hits in the discovery set: SNP-level: $p < 0.05/9$; gene-level: $p < 0.05/5$; pathway-level: $p < 0.05/1$

De novo genome assembly

Next, we used a reference-free *de novo* genome assembly approach to examine associations between presence/absence of genes and the TB phenotype, and to study sequences not present in the reference genome. The list of annotated coding sequences for the 108 assembled genomes in the discovery set contained 3,032 distinct genes that were present in at least one, but not all of the strains (Additional file 8). The permutation analysis revealed six genes that were significantly associated with the TB phenotype. Two of these genes, *Rv3433c* and *nanK*, were validated in the

Table 5.2. Significant genes identified by the *de novo* genome assembly analysis

	Discovery dataset (n=108)			Validation dataset (n=214)		
	Strains with CDS present (N)			Strains with CDS present (N)		
Coding sequence	TBM (n=54)	PTB (n=54)	P-value	TBM (n=52)	PTB (n=162)	P-value^
Bifunctional NAD(P)H-hydrate repair enzyme Nnr (*Rv3433c*)	36	47	0.011	34	144	0.002
TMAO/DMSO reductase	46	53	0.016	49	152	0.592
Antitoxin/MT2731	33	44	0.017	37	130	0.206
NPCBM-associated, NEW3 domain of alpha-galactosidase	15	6	0.022	14	19	0.063
Oxidoreductase molybdopterin binding domain protein	11	3	0.028	7	13	0.197
N-acetylmannosamine kinase (*nanK*)	44	51	0.036	44	157	0.001

NOTE. P-values are based on permutation analysis; bold p-values indicate validated significant enrichment
^ P-value thresholds in the validation set were Bonferroni-corrected for multiple testing by dividing them by the number of top hits in the discovery set: $p < 0.05/6$

validation set: absence of these genes was associated with TBM (Figure 5.1, Table 5.2, Additional file 9). Together with *Rv0218* they bring the total number of genes associated with the TB phenotype to three.

Effect of detected SNPs on protein function and predicted function of phenotype-associated genes

We used published algorithms to predict the effects of identified mutations on protein structure and function. The SNP in *Rv0218*, a protein predicted to have transmembrane helices, likely leads to a decrease of stability of the protein (Table S5.3). For *Rv3433c*, and *nanK* no transmembrane helices or signaling peptides were predicted.

Discussion

To determine whether *M. tuberculosis* genetic variation is associated with the TB disease phenotype, we compared *M. tuberculosis* whole genome sequences from 216 PTB and 106 TBM patients and searched for homoplastic mutations. We identified three genes in *M. tuberculosis* (*Rv0218*, *Rv3433c*, and *nanK*) to be associated with either TBM or PTB. Previous experimental studies have assessed the importance of *Rv0218*. This secretome gene encodes for a protein with multiple predicted transmembrane regions and a C-terminal molybdopterin binding domain that is often found in oxidoreductases and was shown to be essential for *M. tuberculosis in vivo* growth in C57BL/6 J mouse spleen[37]. The SNP in Rv0218 is predicted to decrease the stability of the respective protein. Secretome genes potentially influence pathogen recognition and host-pathogen interaction[38]. If mutations in these genes alter the appearance of the *M. tuberculosis* surface, this could provide a mechanism by which *M. tuberculosis* could evade the immune response and enable dissemination to extrapulmonary sites. Secretome genes are more likely to contain false-positive associations as they are under selective pressure from the immune system and phages[39]. How *Rv3433c* and *nanK* could be related to the TB phenotype is not obvious, although functions have been predicted based on homology detection.

To our knowledge, this is only the second attempt to relate *M. tuberculosis* genetic variation to the TB disease phenotype in humans on a genome-wide scale. The other study, by Saw *et al.* showed large-scale rearrangements, short translocations, inversions, indels and SNPs in eight strains cultured from CSF[33]. Non-synonymous SNPs in eight genes (*embR*, *lppD*, *PE-PGRS10*, *PE-PGRS19*, *PE-PGRS21*, *PE-PGRS49*, *PPE58*, and *Rv0278c*) were found in at least four of the eight CSF-derived strains, and in none of 69 strains isolated from sputum[33]. We did not confirm this in our set of isolates, although *PE-PGRS19* was associated with the TB phenotype in the discovery set. Moreover, we used a two-step approach, based on homoplasy counting as well as allele counting with a correction for phylogenetic bias to find mutations associated with the TB phenotype, and we performed ancestral reconstruction for the most discriminative SNP. Unlike a previous study from Vietnam[32], we found no association between *M. tuberculosis* lineage and TBM. This is no surprise given the genetic diversity even within *M. tuberculosis* lineages[9], and the observed pattern of TBM isolates scattered across the phylogenetic tree.

In concordance with previous findings[9], we found considerable genetic diversity in *M. tuberculosis* in the current study. Two isolates differed on average by 1,000 SNPs, and this did not differ among PTB isolates and among TBM isolates. In addition, we did not observe any clustering, defined as two isolates differing by 12 SNPs or less[35]. The lack

of clustering is probably a result of the low sampling fraction in this urban setting with thousands of incident TB cases each year.

Theoretically, two scenarios could explain the role of *M. tuberculosis* genetic variation in the development of TBM after infection with *M. tuberculosis*. First, upon infection the *M. tuberculosis* strain may carry certain mutations associated with dissemination and penetration of the blood-brain barrier. Second, a subpopulation of bacteria in the lungs of a PTB patient may develop such mutations, though it was recently shown for bacterial meningitis caused by *S. pneumonia* or *N. meningitides* that there is no evidence for differential selection between blood and CSF, and that any mutations between these two niches are likely due to mutation hotspots or forms of diversifying selection common to both niches[40]. However, similar to the findings of Saw *et al.*[33], the genetic variants that we found to be associated with the TB disease phenotype were not exclusive for TBM or PTB, nor were they consistently present in all TBM or PTB strains. Therefore it seems that genetic variants may be part of a complex, multifactorial process leading to this devastating manifestation of TB, in which the human genotype or phenotype equally plays an important role[32,41].

This is the second, and by far largest study using whole genome sequencing to link *M. tuberculosis* genotype to TBM. In this large cohort of well-characterised patients we studied strains from HIV-negative, adult patients to control for the two most important known risk factors for TBM. In addition, both patient groups were similar with regard to gender, ethnicity, and previous episodes of TB. The *de novo* assembly adds to the strengths of this study because it enabled us to examine regions of the genome that do not map to the reference genome, allowing the investigation of associations between genetic variation in these genomic regions and the TB disease phenotype. The homoplasy-based association analysis has proven to be a successful method to detect *M. tuberculosis* loci associated with a certain phenotype (e.g. transmissible vs. non-transmissible, drug-resistant vs. sensitive)[42,43]. The major advantage is that false-positive associations due to genetic relatedness of strains with the same phenotype (i.e. 'phylogenetic bias') are filtered out, thereby increasing statistical power to find true associations. In addition, the ancestral reconstruction in the validation step ruled out the possibility that the significant association for the SNP in *Rv0218* was due to population structure.

The current study has several limitations. Firstly, we only focused on mutations in coding regions of the genome, as they are more likely to have functional consequences, but mutations in non-coding regions could also affect function, for instance by transcriptional and translational regulation of protein-coding sequences[44]. Secondly, the large number of genetic variants increases the risk of finding false-positive

associations, although homoplasy counting enabled us to filter out many of these false-positives. We did not correct for multiple testing in the discovery set, but we used a validation set where we did correct for multiple testing for confirmation. Lastly, whether bacteria developed TBM-associated mutations before or after infecting a patient remains unclear. One way to investigate this is to compare the genomes of strains isolated from sputum and CSF from the same patient. Unfortunately we did not have the availability of paired isolates. Most TBM patients were too ill to expectorate sputum.

Conclusions

We present evidence from a homoplasy-based association analysis that three *M. tuberculosis* genes, including *Rv0218*, a cell wall-associated and/or secretome gene, are associated with the TB disease phenotype. These findings serve as an important step forward in the quest for an improved understanding of the mycobacterial determinants of TB tissue tropism. Functional validation studies are warranted to further explore the effect of mutations in these genes on protein function.

Methods

Patients and isolates

We used *M. tuberculosis* isolates from two established cohorts of Indonesian patients with confirmed TB. The first group consisted of adult patients (≥15 years old) with TBM admitted at Hasan Sadikin Hospital between 2006 and 2013, with *M. tuberculosis* cultured from CSF. The second group was randomly selected from a cohort of culture-positive HIV-negative PTB patients (age ≥ 15 years) from the same setting recruited between 2012 and 2015. All patients were tested for HIV, and those who were HIV-positive were excluded.

Sequencing, alignment, and variant calling

Mycobacterial DNA was extracted from cultures using cetyl trimethylammonium bromide (CTAB) or using UltraClean® Microbial DNA Isolation Kit (MO BIO Laboratories). A single isolate from each patient was selected for sequencing. *M. tuberculosis* DNA was sequenced on an Illumina HiSeq 2000 instrument using 2 x 100 bp paired-end reads at the Beijing Genome Institute in Hong Kong. After sequencing, the raw FASTQ sequence reads were filtered, including removing of adapter sequences, contamination, and low quality reads which have more than 10% N base calls, or where more than 40% of the bases have a quality score ≤ 4. Quality control statistics are shown in Table S5.4. Five TBM strains and four PTB strains were contaminated, based on a low GC-content, and were excluded from further analyses. Sequencing coverage was

determined using the FASTQC quality control tool version 0.10.1. The proportion of bases sequenced with a sequencing error rate of 1% or less per base ranged from 93% to 97% per genome. The average coverage depth for the remaining 322 sequenced strains was 121.1, and the average percentage of bases covered by at least one read was 98.9%.

The sequence reads were aligned to reference strain *M. tuberculosis* H37Rv, accession number NC_000962.3, and variants were called using Breseq software, version 0.27.1[45] using a minimum threshold of 30x coverage. Mutations with low-quality evidence (i.e. possible mixed read alignment) were not included. The Breseq variant call output was converted to a tab-separated file for each sequence using customized Python and R scripts that are available upon request.

Phylogeny construction

A phylogeny was constructed to determine evolutionary relationships of the isolates. We extracted all 29,199 variable positions across the 322 *M. tuberculosis* sequences and concatenated them into a single alignment. Solely for the purpose of creating the phylogenetic tree, SNPs occurring in PE/PPE genes and genes related to mobile elements (genes listed in Table S5.5) were excluded to avoid any concern about inaccuracies in the read alignment in these parts of the genome. In addition, SNPs in an additional 40 genes previously associated with drug resistance[46] were removed to exclude the possibility that homoplasy of drug resistance mutations would significantly affect the phylogeny[47]. After applying these filters to the initial set of 29,199 SNPs, the 28,544 remaining SNPs were used to construct the phylogenetic tree using PhyML, version 3.0[48] using the HKY85 model with four categories for the gamma distribution, and using a hundred bootstraps.

To determine the lineage distribution of the strains and to evaluate whether an association exists between *M. tuberculosis* lineage and TB disease phenotype, we determined the lineage for each of the 322 strains using a 62-SNP barcode[34]. The resulting classification in the main *M. tuberculosis* lineages also served as a quality check for the generated Maximum Likelihood (ML)-phylogenetic tree, as it enabled us to validate that isolates belonging to the same lineage clustered together in the tree. A Chi-square test was used to statistically test the association between *M. tuberculosis* lineage and TB disease phenotype.

Homoplasy-based association test to identify associations between *M. tuberculosis* genotype and TB disease phenotype

We used a two-step approach: in the discovery step we aimed to maximize power by homoplasy counting, without correction for multiple testing. In the subsequent validation step, aimed to distinguish true associations from false positives, we used allele counting with multiple testing correction, and performed ancestral reconstruction to remove possible phylogenetic bias.

To divide the genomes in a discovery and a validation set, we identified isolates in terminal branch pairs, trios and quartets (i.e. separated at a terminal or near terminal branch in the phylogenetic tree) with distinct phenotypes (Figure 5.2). These terminal branch sets (TBSs) together formed the discovery set. These provide the strongest, homoplasy-corrected possible association with the phenotype. We used the remaining genomes to validate the association. For all isolates in the TBSs, we listed the non-synonymous SNPs to create a table with all variable positions in rows, the paired isolates in columns, and the corresponding nucleotide in the cells (Additional file 2). For every SNP, an enrichment score was calculated using the following formula:

$$\log\left(\frac{(\text{number of TBM isolates with SNP}/\text{total number of TBM isolates}) + 0.001}{(\text{number of PTB isolates with SNP}/\text{total number of PTB isolates}) + 0.001}\right)$$

A permutation p-value for each SNP was calculated by randomising the phenotypes over the isolates 1,000 times.

In parallel we grouped SNPs per gene, using the same empirical randomisation strategy to assess association, adjusted for gene length:

$$\log\left(\frac{((\text{number of SNPs in gene in TBM isolates}/\text{gene length})/\text{total number of TBM isolates}) + 0.001)}{((\text{number of SNPs in gene in PTB isolates}/\text{gene length})/\text{total number of PTB isolates}) + 0.001))}\right)$$

Similarly, we grouped genes with ≥1 SNP per *M. tuberculosis* pathway according to PATRIC[49], and calculated association using the aforementioned permutation analysis, adjusted for the number of genes in a particular pathway:

$$\log\left(\frac{((\text{no.of pathway genes with} \geq 1 \text{ SNP in TBM isolates}/\text{unique gene count})/\text{total no.of TBM isolates}) + 0.001}{((\text{no.of pathway genes with} \geq 1 \text{ SNP in PTB isolates}/\text{unique gene count})/\text{total no.of PTB isolates}) + 0.001}\right)$$

Significance of associations was determined by calculating a permutation p-value through randomization of the phenotypes over the isolates 1,000 times. All calculations were performed with customized Perl scripts that are available upon request.

We used the set of 214 strains that were not in TBSs to validate candidate SNPs, genes, and pathways identified in the discovery set, using the same permutation test as described above for the discovery set. We used a p-value threshold of 0.05 for the discovery set. The p-value thresholds in the validation set were Bonferroni-corrected for multiple testing by dividing them by the number of significant (candidate) hits in the discovery set. To correct for potential phylogenetic bias in the validation set, we performed ancestral reconstruction for validated TB phenotype-associated SNPs using FASTML[50] with default parameters, and compared the proportion of TBM vs. PTB isolates prior to (i.e. older than) and after (i.e. younger than) the occurrence of the SNP in the validation set. For each node / leave where the SNP occurred, we calculated the proportion of TBM isolates among the child branches and we calculated the (unweighted) average over all of these nodes and leaves to determine the proportion of TBM isolates after the SNP (Table S5.2). This way, every independent occurrence of the SNP contributes equally to the analysis, regardless of the number of child branches after the SNP, thus correcting for phylogenetic bias. The significance of the difference in proportion was determined by calculating the Z-score for 2 population proportions with accompanying p-value.

PE/PPE genes, a major challenge in the analysis of *M. tuberculosis* whole genome sequences due to the repetitive nature of these sequences, were included in the analysis. TB phenotype-associated SNPs in PE/PPE genes were manually examined to confirm that they did not fall within a repetitive region (for an example please see Figure S5.4).

De novo genome assembly

Sequence reads were *de novo* assembled using SPAdes, version 3.6.1[51] with default parameters. All assemblies were evaluated, focussing on genome size, N50 length, number of contigs and scaffolds, and GC-content. The assembled genomes were annotated using Prokka, version 1.11[52] with default parameters. Completeness and contamination of assemblies were determined with CheckM version 1.0.5[53]. The assembly statistics are shown in Additional file 14. Presence or absence of annotated genes was determined for the 108 assembled genomes in the discovery set. An enrichment score per gene was calculated based on the frequency of occurrence in TBM vs. PTB strains using the following formula:

$$\log\left(\frac{(\text{number of TBM isolates with gene present}/\text{total number of TBM isolates}) + 0.001}{(\text{number of PTB isolates with gene present}/\text{total number of PTB isolates}) + 0.001}\right)$$

Statistical significance was again determined based on permutation by randomizing the phenotypes over the isolates 1,000 times. We repeated this permutation analysis

for the 214 genomes comprising the validation set, using Bonferroni-adjusted p-value thresholds. For the genes with a validated, significant enrichment for TBM or PTB, we confirmed their absence in the respective genomes by mapping the raw sequencing reads for these genomes back to the H37Rv reference sequence of the gene (Figure S5.5), and visualized this with integrative genomics viewer (IGV), version 2.3.32[54].

Prediction of mutation effects

We used two algorithms to predict the effect of the mutations on protein structure and function. I-Mutant version 2.0, which predicts the protein stability change upon single site mutation (http://folding.biofold.org/imutant/i-mutant2.0.html)[55] and PolyPhen-2, which predicts the possible impact of an amino acid substitution on the structure and function of a protein (http://genetics.bwh.harvard.edu/pph2/)[56] to predict the impact of the validated SNPs on protein structure and function. In addition, we used TartgetP (http://www.cbs.dtu.dk/services/TargetP/)[57] to predict the subcellular location of the proteins encoded by the validated genes, and TMHMM (http://www.cbs.dtu.dk/services/TMHMM/)[58] to predict transmembrane helices in these proteins.

Acknowledgements

The authors would like to thank the data management team members for data management, the residents for monitoring patients, professor Jelle Goeman for statistical advice; Jakko van Ingen for fruitful discussions; Jordy Coolen, Maha Farhat, Daniel Garza, Robin van der Lee, and Aldert Zomer for advice on the methodology and bioinformatics; Bruno Andrade for assisting in the *de novo* assembly, and the director of the Hasan Sadikin General Hospital for accommodating the research.

Funding

This study was supported by the Royal Netherlands Academy of Arts and Sciences (KNAW). [09-PD-14 to RvC]; fellowship from the Netherlands Organization for Health Research and Development (ZonMw) and The Netherlands Foundation for Scientific Research [VIDI grant. 017.106.310 to RvC., and VIDI grant 864.14.004 to BED.]; and Radboud University fellowship [to CR]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. This study was also supported by the TANDEM (Tuberculosis and Diabetes Mellitus) Grant of the ECFP7 (European Union's Seventh Framework Programme) under Grant Agreement no. 305279.

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request. Data generated or analysed during this study are included in this published article [and its supplementary information files],

or are available from the corresponding author on reasonable request. The raw sequence files (FASTQ) were archived on the NCBI Sequence Read Archive and are available at: https://www.ncbi.nlm.nih.gov/sra/SRP130118. The individual isolates can be accessed under the following Biosample accession numbers: SAMN08376067-SAMN08376388. The Bioproject accession number is: PRJNA430531. Phylogeny data have been uploaded to TreeBASE (http://purl.org/phylo/treebase/phylows/study/TB2:S22081).

Authors' contributions

CR was responsible for conceptualization, data analysis, funding acquisition, and writing. LC was responsible for laboratory management, project administration, resources, and writing. AvL was responsible for conceptualization and writing. SD was responsible for inclusion of patients in the study and writing. ARG was responsible for patient management and writing. HNG was responsible for conceptualization and methodological guidance. MAH was responsible for conceptualization, methodology, resources, supervision, and writing. BA was responsible for funding acquisition, project administration, resources, and supervision. BED was responsible for conceptualization, data analysis, methodology, supervision, and writing. RvC was responsible for conceptualization, funding acquisition, project administration, supervision, and writing. All authors read and approved the final manuscript.

Ethics approval and consent to participate

All adult patients provided written informed consent; from the age of 15, patients are no longer seen by a paediatrician[59] and parents provided informed consent for patients under 18. The consent procedure was approved by the local Institutional Review Board. The study protocols for the inclusions of patients and for bioanalysis were approved by the ethical committee of the Faculty of Medicine, Universitas Padjadjaran / Hasan Sadikin Hospital, Bandung, Indonesia under ethical registration number 0716040326.

Competing interests

The authors declare that they have no competing interests.

Supplementary figures

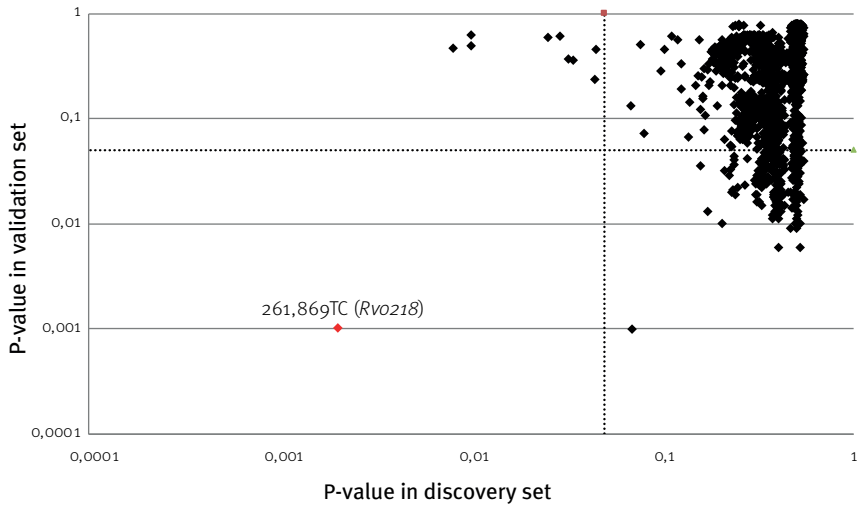


Figure S5.1. Scatterplot showing the p-values of the SNPs found in the discovery and the validation set. P-values in the discovery set are shown on the x-axis; p-values in the validation set are shown on the y-axis. Gene names are shown for SNPs significant in both the discovery and validation set.

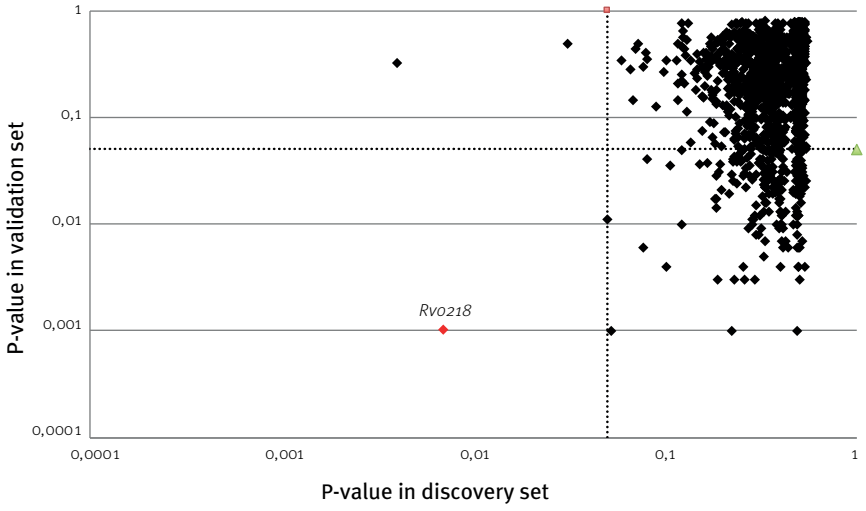


Figure S5.2. Scatterplot showing the p-values of the genes found in the discovery and the validation set. P-values in the discovery set are shown on the x-axis; p-values in the validation set are shown on the y-axis. Names are shown for genes significant in both the discovery and validation set.

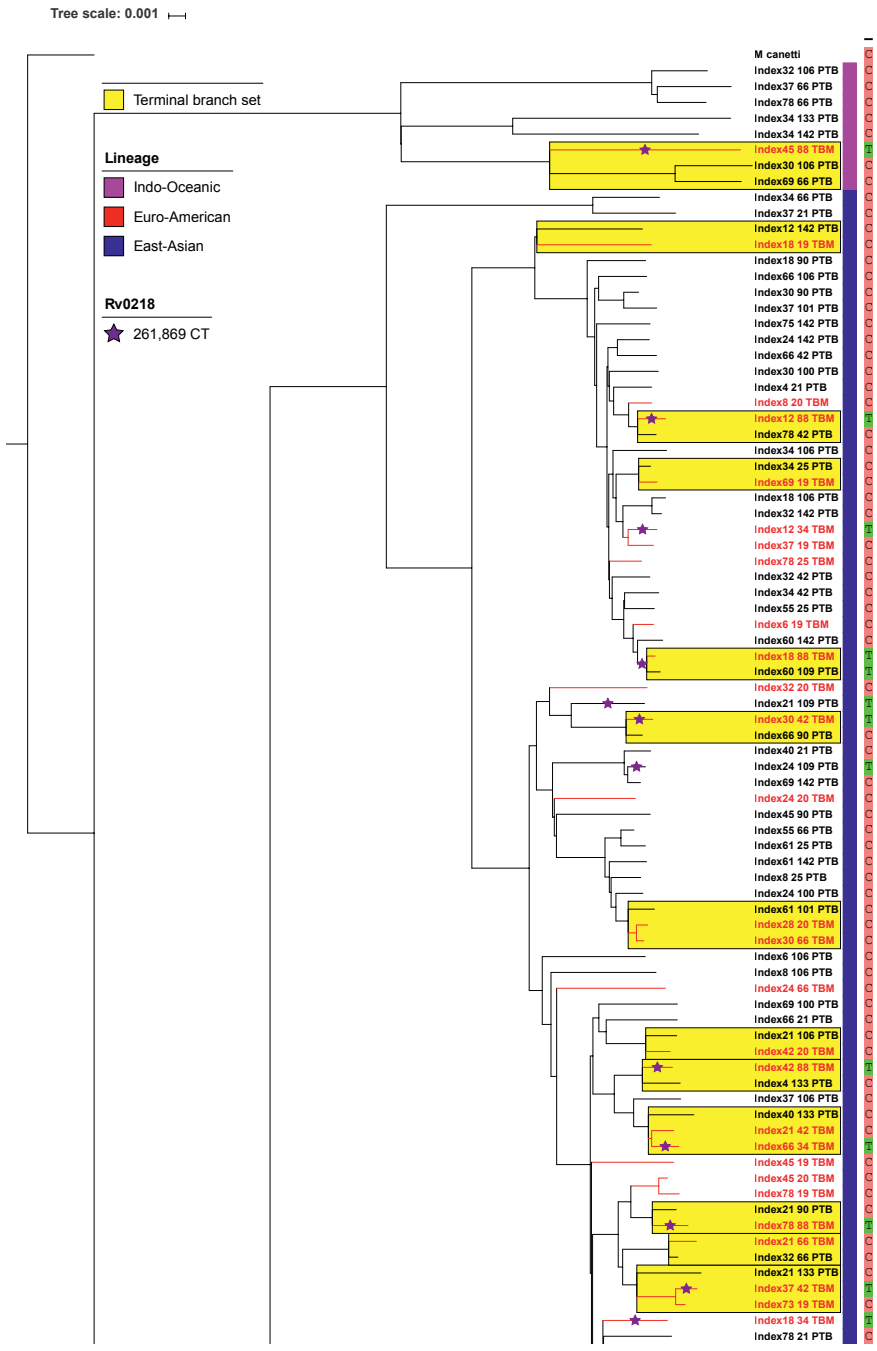

Tree scale: 0.001
Terminal branch set
Lineage
Indo-Oceanic
Euro-American
East-Asian
Rv0218
261,869 CT
M canetti
Index32 106 PTB
Index37 66 PTB
Index78 66 PTB
Index34 133 PTB
Index34 142 PTB
Index45 88 TBM
Index30 106 PTB
Index69 66 PTB
Index34 66 PTB
Index37 21 PTB
Index12 142 PTB
Index18 19 TBM
Index18 90 PTB
Index66 106 PTB
Index30 90 PTB
Index37 101 PTB
Index75 142 PTB
Index24 142 PTB
Index66 42 PTB
Index30 100 PTB
Index4 21 PTB
Index8 20 TBM
Index12 88 TBM
Index78 42 PTB
Index34 106 PTB
Index34 25 PTB
Index69 19 TBM
Index18 106 PTB
Index32 142 PTB
Index12 34 TBM
Index37 19 TBM
Index78 25 TBM
Index32 42 PTB
Index34 42 PTB
Index55 25 PTB
Index6 19 TBM
Index60 142 PTB
Index18 88 TBM
Index60 109 PTB
Index32 20 TBM
Index21 109 PTB
Index30 42 TBM
Index66 90 PTB
Index40 21 PTB
Index24 109 PTB
Index69 142 PTB
Index24 20 TBM
Index45 90 PTB
Index55 66 PTB
Index61 25 PTB
Index61 142 PTB
Index8 25 PTB
Index24 100 PTB
Index61 101 PTB
Index28 20 TBM
Index30 66 TBM
Index6 106 PTB
Index8 106 PTB
Index24 66 TBM
Index69 100 PTB
Index66 21 PTB
Index21 106 PTB
Index42 20 TBM
Index42 88 TBM
Index4 133 PTB
Index37 106 PTB
Index40 133 PTB
Index21 42 TBM
Index66 34 TBM
Index45 19 TBM
Index45 20 TBM
Index78 19 TBM
Index21 90 PTB
Index78 88 TBM
Index21 66 TBM
Index32 66 PTB
Index21 133 PTB
Index37 42 TBM
Index73 19 TBM
Index18 34 TBM
Index78 21 PTB

Figure S5.3.

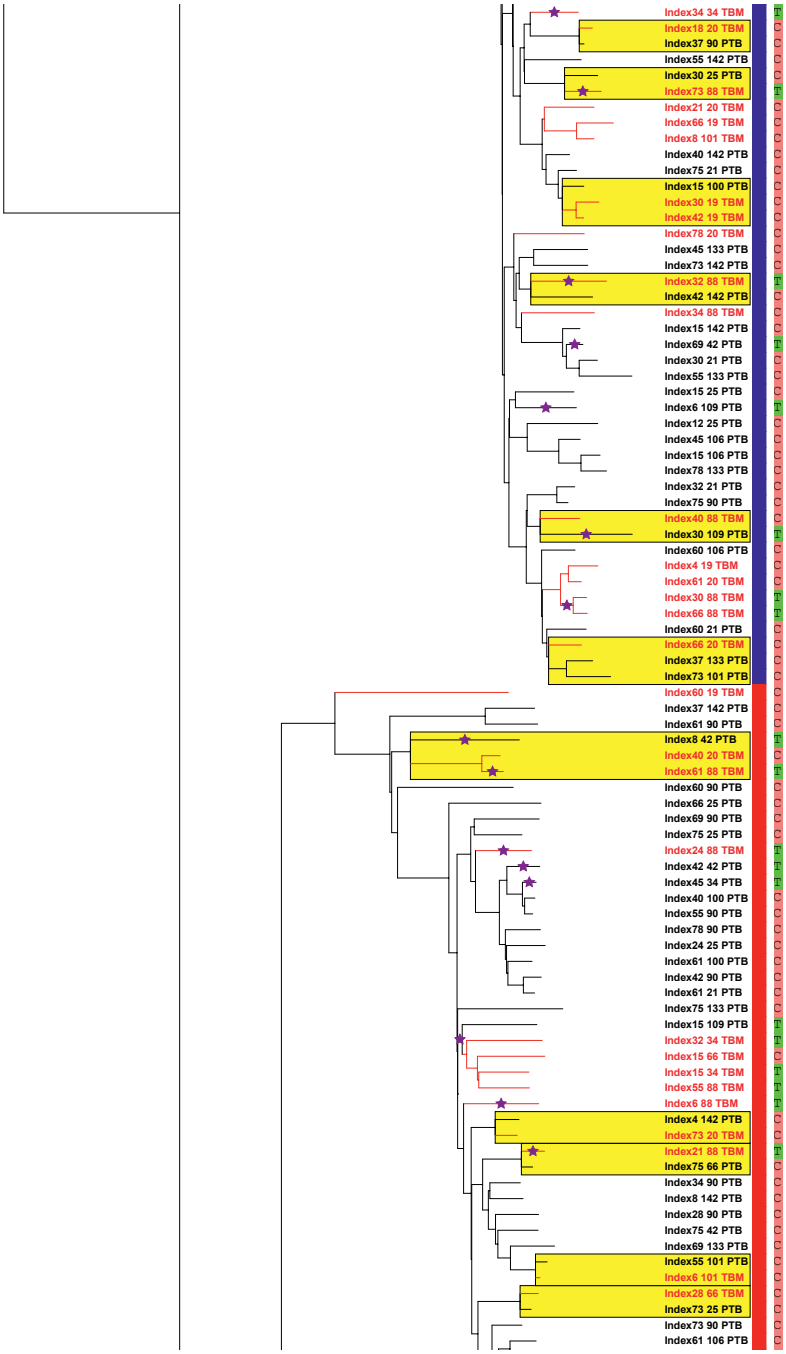


Figure S5.3. Continued.

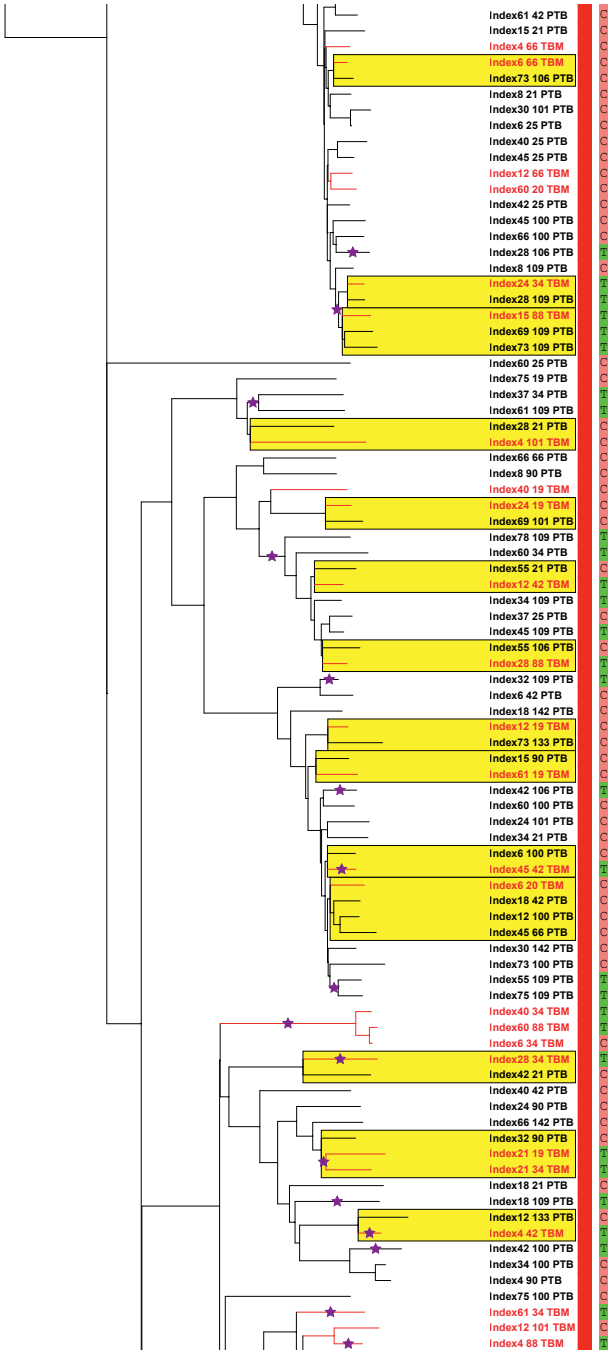


Figure S5.3. Continued.

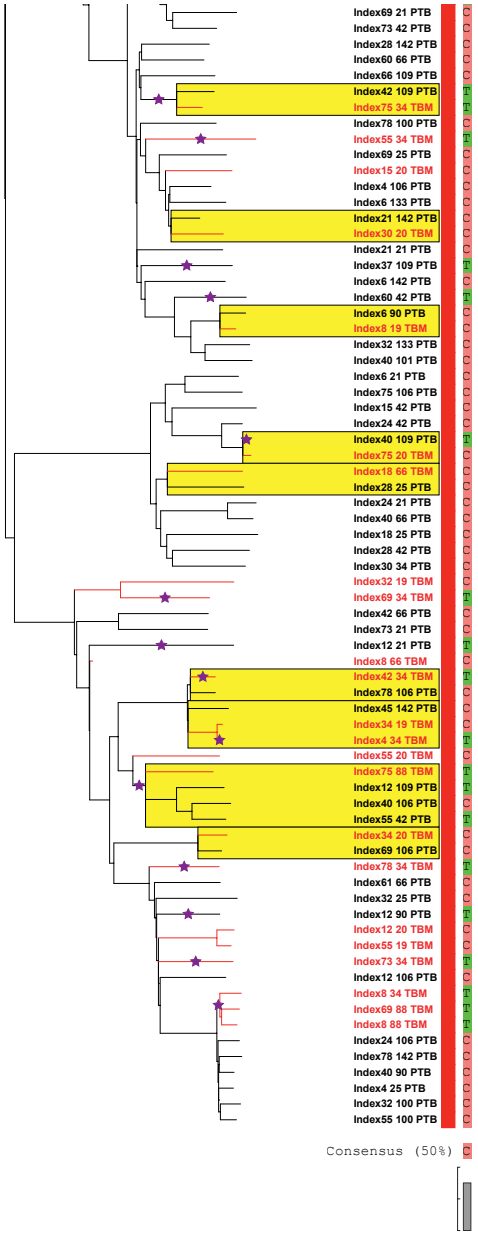


Figure S5.3. Phylogenetic tree of 322 *M. tuberculosis* strains isolated from TBM (red branches and leaves) and PTB (black branches and leaves) patients. The highlighted branches indicate the 108 strains in 47 terminal branch sets, together comprising the discovery set. The purple stars indicate the origin of the SNP in *Rv0218* according to the ancestral reconstruction. The nucleotide for SNP position 261,869 is indicated next to the leaf labels.

Figure S5.4. Diagram demonstrating breseq calling a SNP in the *PE-PGRS1* gene. Displayed are 60 Illumina sequencing reads mapping to the H37Rv reference genome (shown at the top and bottom). Visual inspection of the SNP confirms that it does not occur in a region containing uniformly lower base quality scores.

Figure S5.5. Read alignment demonstrating the absence of *nanK*. Displayed is the alignment of the raw sequencing reads against the H37Rv *nanK* gene. No reads are mapping to this gene, showing that it is absent in the sequenced genome.

Supplementary tables

Table S5.1. Description of baseline characteristics for PTB and TBM patients

Characteristic	TBM patients (N=106)	PTB patients (N=216)
Male gender – N (%)	58 (54.7)	111 (51.4)
Age (yrs) – mean (SD)	28.0 (8.9)	39.1 (14.8)
Ethnicity – N (%)		
Sundanese	45 (93.7)	187 (87.4)
Javanese	3 (6.3)	21 (9.8)
Sumatra	0 (0)	3 (1.4)
Other	0 (0)	3 (1.4)
Years after admission – median (IQR)	6 (4-8)	1 (1-2)
History of TB treatment - N (%)	13 (13.1)	37 (17.2)

IQR, interquartile range; SD, standard deviation.
* Data were missing for history of TB treatment (TBM, n=7; PTB, n=1); ethnicity (TBM, n=58; PTB = 2).

Table S5.2. Ancestral reconstruction of SNP 261869TC in *Rv0218*

Node/leave where SNP occurred	Number of child branches	Number of TBM branches among children	Percentage TBM branches among children
Index12_34_TBM	1	1	100
Index21_109_PTB	1	0	0
Index24_109_PTB	1	0	0
Index18_34_TBM	1	1	100
Index34_34_TBM	1	1	100
Index69_42_PTB	1	0	0
Index6_109_PTB	1	0	0
Index30_88_TBM|Index66_88_TBM	2	2	100
Index24_88_TBM	1	1	100
Index42_42_PTB	1	0	0
Index45_34_PTB	1	0	0
Index15_109_PTB|Index32_34_TBM|Index15_66_TBM|Index15_34_TBM|Index55_88_TBM	4	3	75
Index6_88_TBM	1	1	100
Index28_106_PTB	1	0	0
Index37_34_PTB|Index61_109_PTB	2	0	0
Index78_109_PTB|Index60_34_PTB|Index34_109_PTB|Index37_25_PTB|Index45_109_PTB	4	0	0
Index32_109_PTB	1	0	0
Index42_106_PTB	1	0	0
Index55_109_PTB|Index75_109_PTB	2	0	0
Index40_34_TBM|Index60_88_TBM|Index6_34_TBM	2	2	100
Index18_109_PTB	1	0	0
Index42_100_PTB	1	0	0
Index61_34_TBM	1	1	100
Index4_88_TBM	1	1	100
Index55_34_TBM	1	1	100
Index37_109_PTB	1	0	0
Index60_42_PTB	1	0	0
Index69_34_TBM	1	1	100
Index12_21_PTB	1	0	0
Index78_34_TBM	1	1	100
Index12_90_PTB	1	0	0
Index73_34_TBM	1	1	100
Index8_34_TBM|Index69_88_TBM|Index8_88_TBM	3	3	100
Sum	45	21	1475
Average	1,36	0,64	44,70

Listed are the internal nodes and leaves where the SNP in *Rv0218* occurred according to the ancestral reconstruction of the SNP

Table S5.3. Protein prediction for genomic sites associated with the TB disease phenotype

				Protein Prediction		Predicted localization	Predicted number of transmembrane helices
	Genomic position	Nucleotide change	Amino acid change	I-Mutant #	PolyPhen &	TargetP ^	TMHMM *
SNP-level	Rv0218						
	261,869	C=›T	R316C	Decrease of Stability	Benign	Other	5
Gene-level	Rv3433c					Other	0
	nanK					Other	0

I-mutant predicts free energy changes of protein stability upon a point mutation under different conditions.
& PolyPhen predicts the possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations.
^ TargetP predicts the subcellular location of proteins based on the predicted presence of any N-terminal signal peptides.
* TMHMM predicts transmembrane helices in proteins.

Table S5.4. Description of sequencing quality control parameters and statistics

Study number	Clean reads	Q20%	Total bases	Mean coverage	Median coverage	Percentage bases >1
100865	5393826	94,86	496846191	112,60	117	99,00
100866	2806938	95,62	258749297	58,70	61	99,10
100870	3469296	94,54	320142664	72,60	75	99,20
100874	8497168	95,62	783313115	177,60	184	99,70
100876	6903240	95,75	636549074	144,30	150	99,50
100878	8828962	95,59	814420068	184,60	191	99,50
100879	4471336	95,48	412040142	93,40	97	99,00
100883	11819814	95,73	1089505119	247,00	255	99,20
100884	3347584	95,08	309679647	70,20	73	99,50
100886	7686658	95,78	709079468	160,70	167	99,20
100887	4668126	95,28	430856934	97,70	101	99,40
100890	8814420	95,67	814649331	184,70	192	99,20
100891	5695014	95,52	524832508	119,00	123	99,10
100892	2735174	95,89	251826716	57,10	59	99,00
100894	6729510	95,83	620328102	140,60	145	99,50
100895	9433682	95,74	871716597	197,60	205	99,20
100896	8746936	95,72	806976809	182,90	190	99,80
100901	4186882	95,06	386788589	87,70	91	99,20
100904	8058216	95,56	744905643	168,90	175	99,40
100906	5458790	95,47	455994928	103,40	107	99,30
100907	9489138	95,51	876888230	198,80	207	98,70
100912	4186690	94,61	383674366	87,00	89	99,20
100913	6683380	94,67	616012210	139,60	146	98,80
100917	5527220	95,57	508711705	115,30	119	99,90
100918	6806576	94,53	628084709	142,40	148	98,80
100921	3123650	95,05	288755589	65,50	68	99,50
100923	10913652	94,92	1007150244	228,30	238	99,40
100925	8613960	94,75	795469596	180,30	188	98,80
100926	9149536	95,67	832817109	188,80	195	99,30
100927	7666522	95,70	707747568	160,40	165	99,90
100928	5247660	95,03	483371397	109,60	114	98,90
100929	6371238	95,54	587360512	133,10	138	99,00
100930	7564418	95,32	698275127	158,30	163	99,50
100931	9221108	94,85	852549574	193,30	202	99,20
100932	6497346	94,98	599473298	135,90	142	99,20
100933	6973496	94,98	642622186	145,70	152	99,40
100935	2267830	94,46	205539485	46,60	45	99,30
100941	2677256	95,41	244781931	55,50	54	99,10
100943	3170650	94,63	292561071	66,30	69	98,90
100946	5172654	94,71	477425025	108,20	113	98,90
100947	8855840	95,02	816620961	185,10	193	99,20
100951	2767780	95,39	255056580	57,80	57	99,10

Table S5.4. Continued

Study number	Clean reads	Q20%	Total bases	Mean coverage	Median coverage	Percentage bases >1
100960	4637948	95,43	425928568	96,60	99	99,10
100962	4310596	95,63	392051064	88,90	91	99,20
100963	1974408	94,21	182390070	41,30	43	98,80
100964	2262778	94,91	208683478	47,30	48	99,30
100967	2711596	95,22	248967022	56,40	58	99,40
100968	5123952	95,30	471987706	107,00	110	99,00
100972	7477410	95,21	689739641	156,40	161	99,20
100975	7998754	95,34	738615844	167,40	173	99,50
100976	3638594	94,75	335773295	76,10	80	98,80
100983	3669988	94,48	339045575	76,90	80	98,80
100984	6956916	95,35	641494402	145,40	145	99,40
100987	2629362	95,11	243150452	55,10	57	99,10
100989	11563038	95,77	1065686827	241,60	249	99,10
100991	6316160	94,96	582725768	132,10	138	98,90
100992	2859276	94,60	263619555	59,80	62	98,50
100993	8231110	95,79	758381553	171,90	178	99,20
100994	5536364	94,88	510337890	115,70	121	98,80
100996	7075788	95,07	651919072	147,80	153	99,20
100997	2132396	94,05	197331995	44,70	47	98,40
100999	6085482	95,58	556429986	126,10	129	99,60
900077	4657458	95,37	430691102	97,63	103	98,40
900083	6691044	95,56	617998403	140,09	147	98,20
900084	3248492	95,18	299922817	67,99	71	98,30
900089	3523302	95,20	324308817	73,51	76	99,30
900094	3387992	95,47	312171242	70,76	73	99,20
900095	4466868	94,91	411820079	93,40	94	99,10
900104	10781194	95,73	996184664	225,80	234	99,10
900110	6960366	96,63	642971426	145,75	150	99,80
900115	5065176	95,64	466750089	105,80	110	98,80
900126	4039714	95,29	373033196	84,56	89	98,00
900130	8172166	94,80	755611366	171,30	175	99,00
900135	6004426	95,27	555768696	126,00	130	99,30
900138	7583758	95,44	700928339	158,90	165	99,20
900144	4087780	95,28	376924265	85,44	88	99,10
900150	6564174	95,46	604493321	137,03	143	98,50
900157	13620988	95,57	1261781723	286,02	296	99,00
900159	5053700	95,72	465523074	105,50	109	99,00
900160	791534	94,83	72810185	16,50	17	98,90
900161	6767554	95,40	622483695	141,10	146	99,50
900164	5978184	94,93	551296262	125,00	130	99,00
900166	8463916	94,68	783004573	177,50	184	98,70
900169	3531234	95,09	324288155	73,50	73	99,20

Table S5.4. Continued

Study number	Clean reads	Q20%	Total bases	Mean coverage	Median coverage	Percentage bases >1
900173	4865874	94,96	448459089	101,70	105	99,10
900174	6456632	94,66	596399079	135,20	140	98,80
900176	5891534	94,74	545263881	123,60	129	98,60
900177	8367724	95,47	769016835	174,30	180	99,10
900178	5884278	94,75	544283139	123,40	128	98,80
900179	3757352	94,53	346790136	78,60	82	98,50
900181	3658838	94,80	338192325	76,66	80	98,30
900183	4112656	94,50	379449608	86,00	89	99,50
900190	5386042	95,48	496968260	112,65	117	99,60
900192	4411450	96,52	407585797	92,39	95	99,60
900199	9405206	95,68	869958388	197,20	204	99,30
900201	2187488	95,12	201682848	45,70	47	99,10
900202	8860806	95,83	818983627	185,70	192	99,20
900204	4196972	94,37	387936608	87,90	92	98,50
900206	6629018	94,70	613385999	139,00	145	98,60
900211	8882108	95,37	819024355	185,70	191	99,10
900212	8926712	95,66	822954141	186,60	193	99,00
900215	6848790	95,51	630580535	142,90	147	99,10
900221	2812730	95,25	260233372	58,99	62	98,30
900223	3083040	95,31	284289254	64,44	67	99,10
900225	4983966	95,07	460812198	104,50	108	99,50
900229	6520394	94,75	601645866	136,40	142	98,80
900236	12376840	96,27	1127089142	255,50	261	99,70
900238	9682506	95,50	895230730	202,93	213	98,30
900239	6366596	96,72	587033189	133,07	137	99,30
900241	5696554	96,58	527413573	119,55	123	99,40
900244	7550754	95,37	696142653	157,80	163	99,10
900251	5570528	95,79	513273278	116,40	120	99,30
900319	4327172	95,17	393895196	89,29	88	99,20
900343	4226624	95,62	390290590	88,47	93	97,90
900344	7305994	94,54	673506466	152,70	159	98,80
900365	3633014	95,68	336204924	76,21	80	98,40
900374	7682274	95,54	709433935	160,81	169	98,10
900387	4761330	95,25	438801669	99,50	99	99,30
900394	5895104	95,17	545589186	123,70	128	99,10
900397	6781232	95,65	626276164	141,96	149	98,50
900400	7612354	95,39	703612915	159,49	167	98,20
900406	7457526	95,41	685335643	155,40	159	99,20
900413	8074132	94,82	747137218	169,40	177	98,70
900416	5346728	94,62	494699921	112,10	117	98,70
900417	3282346	94,57	302572231	68,60	72	98,30
900421	6792232	94,41	628043107	142,40	149	98,60

Table S5.4. Continued

Study number	Clean reads	Q20%	Total bases	Mean coverage	Median coverage	Percentage bases >1
900428	6919752	94,64	640814487	145,30	153	98,00
900430	4576412	95,83	421902959	95,60	99	98,80
900437	10100234	94,68	932826377	211,50	220	98,40
900439	1719270	95,22	158735817	35,98	38	97,90
900443	6746410	95,67	622308095	141,10	146	99,30
900446	3568466	95,49	328026972	74,40	76	99,10
900459	3350726	94,33	309370206	70,10	73	98,70
900464	10104002	95,89	931474859	211,20	216	99,10
900469	5572320	95,63	514346967	116,60	120	99,60
900477	10793432	94,58	996500701	225,90	236	98,60
900489	5893196	95,85	543161841	123,10	128	98,80
900490	5546714	95,61	512275952	116,12	122	98,20
900517	5739918	94,60	531133943	120,40	126	98,60
900573	7510088	95,56	693375210	157,17	165	98,60
900574	7482078	95,25	688955525	156,17	162	99,10
900576	9422310	95,75	868555554	196,90	203	98,80
900580	2745100	95,70	232761189	52,80	54	99,00
900586	2581044	95,48	238334139	54,03	57	98,00
900601	9062300	95,53	836943089	189,72	199	98,20
900603	6675866	95,62	616819650	139,80	145	99,20
900609	4064352	94,30	375056975	85,00	89	98,30
900611	6260624	95,40	577124778	130,80	135	99,30
900612	5889964	95,58	542995499	123,10	127	99,00
900618	5136100	95,46	473770925	107,40	110	99,10
900637	2483164	95,21	229386411	52,00	54	99,20
900639	8184572	95,33	757111404	171,60	177	99,30
900654	4897320	95,37	450923632	102,20	105	99,30
900655	3036462	94,58	280487243	63,60	67	98,70
900666	5199786	95,85	478841323	108,50	112	99,00
900677	3880296	94,56	358103680	81,20	85	98,60
900680	5220468	94,67	481344407	109,10	114	98,70
900703	2435754	94,78	225517592	51,10	53	98,60
900706	4981156	95,53	461099564	104,52	109	98,20
900711	4677424	94,92	430474568	97,60	99	99,30
900716	10183442	94,59	938814696	212,80	222	98,80
900731	5091556	95,37	470191118	106,58	112	98,00
900746	6474714	95,36	598086075	135,57	142	98,10
900748	6584354	95,10	607985996	137,82	144	98,40
900751	3003642	95,37	276844453	62,75	65	99,00
900824	6534112	96,35	578749267	131,20	131	99,40
900845	3141548	94,36	290278154	65,80	69	98,60
900849	4992552	95,91	460361996	104,40	108	99,30

Table S5.4. Continued

Study number	Clean reads	Q20%	Total bases	Mean coverage	Median coverage	Percentage bases >1
1001002	8421510	94,85	779007911	176,60	184	99,20
1001003	3107372	95,52	286077988	64,90	67	99,10
1001009	4275054	95,13	394554574	89,40	93	99,30
1001011	5828312	94,85	537712611	121,90	127	99,30
1001012	5244854	95,36	483589099	109,60	113	99,60
1001013	8923730	95,41	822488632	186,40	193	99,70
1100213	4266778	94,92	394750818	89,48	93	98,90
1100217	5838902	95,61	539199148	122,22	125	98,80
1100231	4172814	95,42	384740674	87,21	90	98,90
1100232	6931214	95,41	637630282	144,54	149	99,40
1100233	6438244	96,59	591744287	134,14	136	99,60
1100234	2448826	95,04	226034251	51,24	54	97,90
1100238	5349728	95,45	493746431	111,92	115	99,20
1100246	4241148	94,92	392868330	89,05	92	99,00
1100250	5686092	95,38	525154268	119,04	125	98,10
1100254	691128	95,19	63775796	14,46	15	97,50
1100256	6861744	95,09	635722606	144,10	151	98,50
1100262	1867970	92,89	172516103	39,11	41	98,90
1100264	8528690	95,17	788677462	178,78	185	99,20
1100269	5137284	95,01	474920172	107,65	112	99,10
1100270	6399968	95,37	589507896	133,63	138	99,00
1100273	6762182	95,34	625911758	141,88	149	98,30
1100280	4627628	95,43	427362057	96,87	101	98,40
1100281	6110340	95,50	564048380	127,86	132	99,20
1100283	4638990	95,44	429132510	97,28	102	98,50
1100291	5293572	96,58	488900426	110,82	114	99,70
1100294	7450520	95,38	688696972	156,11	160	99,40
1100300	3092126	95,00	286048335	64,84	68	98,30
1100306	5766562	95,54	532393055	120,68	127	98,30
1100307	5090484	94,61	470152143	106,57	111	98,00
1100315	4590438	95,26	423829736	96,07	101	98,00
1100316	4639218	95,00	429207921	97,29	102	98,20
1100320	4273028	95,14	394124860	89,34	93	98,80
1100331	2893908	95,10	267254684	60,58	63	98,20
1100332	3547096	95,38	327284397	74,19	76	99,10
1100333	7640658	95,53	702530395	159,25	164	99,40
1100335	4719994	95,23	436845013	99,02	104	98,10
1100340	10582420	95,47	977856076	221,66	232	98,60
1100345	5664472	95,22	523635379	118,70	124	98,50
1100348	9478592	96,71	874252406	198,17	204	99,20
1100353	7157040	96,90	662329285	150,14	155	99,60
1100367	4378604	96,73	403022547	91,36	94	99,40

Table S5.4. Continued

Study number	Clean reads	Q20%	Total bases	Mean coverage	Median coverage	Percentage bases >1
1100369	3614860	95,09	334572793	75,84	79	98,40
1100373	7817932	96,89	721194175	163,48	169	99,30
1100374	5687648	95,12	524404441	118,87	123	99,30
1100396	4464604	96,87	412472505	93,50	96	99,60
1100397	7952458	95,27	734301221	166,45	172	99,10
1100400	4156544	94,85	384729079	87,21	91	98,50
1100421	8161636	96,93	744477220	168,76	174	99,30
1100607	6876462	95,38	632158432	143,30	148	99,40
1100707	4391668	95,52	405964237	92,02	96	98,90
1100708	5804762	95,37	535042353	121,28	126	98,90
1100716	8686226	95,58	800338659	181,42	188	98,80
1100730	8065486	96,76	746592857	169,24	174	99,70
1100738	5401532	95,40	500478912	113,45	118	98,30
1100746	3480194	95,27	321712017	72,93	75	98,80
1100757	6216226	95,37	572081247	129,68	130	98,90
1100760	3239756	95,38	298617464	67,69	71	98,50
1100763	5502764	95,21	506670287	114,85	117	98,70
1100764	5650758	96,92	521338278	118,18	122	99,30
1100773	5049028	95,31	466844260	105,82	109	99,10
1100774	1121332	95,44	103410546	23,44	25	98,50
1100786	5043614	95,42	463788799	105,13	109	99,10
1100797	6867872	95,63	631656757	143,18	147	99,30
1100903	3940012	95,42	363569155	82,41	85	98,60
1100907	6662502	96,74	616858915	139,83	144	99,40
1100908	4866202	95,41	447414934	101,42	101	99,00
1100909	4321576	95,18	398668836	90,37	94	98,70
1100917	1474708	92,91	136437753	30,93	32	98,40
1100918	8666842	95,55	800281865	181,41	190	98,40
1100921	5637528	96,81	521923240	118,31	122	99,60
1100929	4545226	95,44	419603339	95,12	98	99,20
1100930	3966012	95,49	367144208	83,22	87	98,30
1100931	4305168	95,24	398774103	90,39	94	99,40
1100933	1616234	92,92	149443084	33,88	35	98,40
1100936	4857908	95,07	448087877	101,57	105	98,60
1100937	4023822	95,29	371528087	84,22	86	98,90
1100942	3877632	95,55	357512894	81,04	84	98,70
1100943	4256638	95,07	391629990	88,77	92	99,00
1100945	3707196	95,44	342392133	77,61	81	98,50
1100969	9138662	95,44	843500045	191,20	197	99,20
1100977	5068004	95,36	467957429	106,08	111	98,20
1100981	5409442	95,53	499060319	113,13	117	98,70
1100984	5795768	95,60	535918638	121,48	126	99,00

Table S5.4. Continued

Study number	Clean reads	Q20%	Total bases	Mean coverage	Median coverage	Percentage bases >1
1100987	2431690	96,80	392597634	88,99	92	99,00
1100995	4298726	95,24	397765175	90,16	93	99,10
1101003	8605192	95,75	793618170	179,90	186	99,20
1101004	5345552	95,18	493292026	111,82	116	99,20
1101008	6612834	96,66	610379624	138,36	142	99,30
1101011	5311910	95,44	490949034	111,29	115	98,90
1101012	4300000	95,19	397980899	90,21	93	98,90
1101015	8385930	95,79	774182342	175,49	179	98,90
1101018	2563844	95,70	237182128	53,76	56	98,80
1101025	9120244	95,95	843597287	191,23	198	98,90
1101028	1896992	95,80	175609777	39,81	41	98,80
1101037	5087396	95,77	471034289	106,77	111	99,20
1101047	7861958	95,42	727939191	165,01	173	98,30
1101054	7246156	95,44	670754167	152,05	159	98,40
1101061	8923668	96,99	823426144	186,65	193	99,30
1101064	4408312	95,68	406658628	92,18	95	98,50
1101067	8193818	96,70	756937816	171,58	177	99,60
1101074	3883520	95,62	358280741	81,21	85	98,50
1101077	3228516	95,68	298974274	67,77	70	98,70
1101080	4911704	95,80	453246251	102,74	107	98,70
1101084	5692096	95,77	525457900	119,11	124	98,60
1101085	4265798	94,95	394010661	89,31	93	98,80
1101088	3440936	95,98	318639444	72,23	75	98,70
1101089	6002548	96,96	556190844	126,08	130	99,60
1101253	3642584	95,74	336144308	76,20	79	98,80
1101292	6082440	95,86	560795868	127,12	132	98,80
1101303	9312806	95,99	861349125	195,25	202	99,20
1101304	2045310	93,60	188641178	42,76	44	98,50
1101316	10031646	95,64	923436067	209,32	214	99,00
1101317	4281276	94,97	393297145	89,15	92	99,00
1101318	8311646	95,94	767350601	173,94	180	99,10
1101319	5606644	95,88	518165835	117,46	122	99,00
1101321	13908338	95,81	1283563813	290,96	303	99,20
1101322	5388084	95,89	496954407	112,65	117	98,70
1101323	4780336	95,88	441188685	100,01	103	98,80
1101324	10556482	96,05	974353093	220,87	228	99,20
1101329	5265746	95,54	485933498	110,15	114	98,60
1101337	9036512	95,73	834483421	189,16	196	98,90
1101340	5004558	95,76	456531484	103,49	102	99,00
1101342	8516402	95,68	787938504	178,61	185	99,00
1101344	5945098	95,92	548757919	124,39	129	98,90
1101346	5892778	95,82	541723597	122,80	127	98,90

Table S5.4. Continued

Study number	Clean reads	Q20%	Total bases	Mean coverage	Median coverage	Percentage bases >1
1101348	7181066	95,79	662390301	150,15	156	98,80
1101350	5585560	95,70	517096265	117,21	122	98,80
1101355	5142926	95,59	475691952	107,83	112	98,70
1101359	8357208	95,68	769440613	174,42	180	99,00
1101360	6005832	95,75	554392328	125,67	130	99,10
1101367	2839034	95,73	262799618	59,57	62	98,80
1101368	5295318	95,94	488701549	110,78	115	98,40
1101382	2494612	95,72	229415793	52,00	54	98,50
1101386	9713036	95,80	896253841	203,16	211	99,10
1101554	8768224	95,99	809038817	183,39	190	99,00
1101556	2695552	93,65	248575069	56,35	58	98,30
1101557	5974612	95,54	552272734	125,19	130	98,80
1101572	9706246	95,94	895520927	203,00	211	98,80
1101576	4786726	95,86	441905771	100,17	104	98,70
1101580	7791782	95,82	718652884	162,90	169	98,90
1101600	3010808	95,87	278389255	63,10	65	98,70
1101602	3259454	95,92	299698944	67,94	71	98,80
1101603	7436708	96,04	686272907	155,56	161	98,90
1101604	6303476	95,18	581855643	131,89	136	98,90
1101610	5124546	96,59	472976032	107,21	111	99,30
1101623	5027232	95,51	463403211	105,04	108	98,70
1101628	5667660	95,74	524058595	118,79	123	99,10
1101652	7220436	96,62	666135610	151,00	156	99,30
1101933	4298652	95,27	397732878	90,16	93	99,10
1101942	4273636	96,65	395475578	89,65	92	99,60
1101947	5441952	96,70	499732433	113,28	115	99,30
1101984	4843912	96,79	446859913	101,29	105	99,20
1101997	5845746	96,62	538540400	122,08	126	99,80
Samples excluded from further analysis						
100973	10916526	96,14	1042435	0,12	1	2,40
900089	6737346	96.70	289038979	32,76	34	99,25
900132	3437888	97.96	32333418	7,33	8	96,90
900248	1760232	96.56	NA	NA	NA	NA
900558	1303536	97.11	NA	NA	NA	NA
900625	5877240	95.53	1290461	0,29	1	10,70
1100343	8561266	95.67	6480416	1,47	2	60,60
1100772	1340688	94.87	NA	NA	NA	NA
1101001	8104650	96.05	275263806	62,30	65	98,30

Displayed are various measures of sequencing quality, used for the sequencing quality control check.

Table S5.5. PE / PPE genes and drug resistance genes excluded for the phylogeny construction.

PE/PPE genes and genes in repetitive regions						Known drug resistance genes
Rv0031	Rv0922	Rv1575	Rv2107	Rv2741	Rv3381c	
Rv0096	Rv0977	Rv1576c	Rv2108	Rv2768c	Rv3386	
Rv0109	Rv0978c	Rv1577c	Rv2123	Rv2769c	Rv3387	
Rv0124	Rv0980c	Rv1578c	Rv2126c	Rv2770c	Rv3388	accD6
Rv0151c	Rv1034c	Rv1579c	Rv2162c	Rv2791c	Rv3425	ahpC
Rv0152c	Rv1035c	Rv1580c	Rv2167c	Rv2810c	Rv3426	efpA
Rv0159c	Rv1036c	Rv1581c	Rv2168c	Rv2812	Rv3427c	embA
Rv0160c	Rv1039c	Rv1582c	Rv2177c	Rv2814c	Rv3428c	embB
Rv0256c	Rv1040c	Rv1583c	Rv2278	Rv2815c	Rv3429	embC
Rv0278c	Rv1041c	Rv1584c	Rv2279	Rv2853	Rv3430c	embR
Rv0279c	Rv1042c	Rv1585c	Rv2328	Rv2885c	Rv3474	ethA
Rv0280	Rv1047	Rv1586c	Rv2340c	Rv2892c	Rv3475	fabD
Rv0285	Rv1054	Rv1646	Rv2352c	Rv2943	Rv3477	fadE24
Rv0286	Rv1067c	Rv1651c	Rv2353c	Rv2943A	Rv3478	fbpC
Rv0297	Rv1068c	Rv1705c	Rv2354	Rv2944	Rv3507	furA
Rv0304c	Rv1087	Rv1706c	Rv2355	Rv2961	Rv3508	gid
Rv0305c	Rv1088	Rv1753c	Rv2356c	Rv2978c	Rv3511	gyrA
Rv0335c	Rv1089	Rv1756c	Rv2371	Rv3018A	Rv3512	gyrB
Rv0354c	Rv1091	Rv1757c	Rv2396	Rv3018c	Rv3514	inhA
Rv0355c	Rv1135c	Rv1763	Rv2408	Rv3021c	Rv3532	iniA
Rv0387c	Rv1149	Rv1764	Rv2424c	Rv3022A	Rv3533c	iniB
Rv0388c	Rv1168c	Rv1765A	Rv2430c	Rv3022c	Rv3539	iniC
Rv0442c	Rv1169c	Rv1768	Rv2431c	Rv3023c	Rv3558	kasA
Rv0453	Rv1172c	Rv1787	Rv2479c	Rv3115	Rv3590c	katG
Rv0532	Rv1195	Rv1788	Rv2480c	Rv3125c	Rv3595c	fabG1
Rv0578c	Rv1196	Rv1789	Rv2487c	Rv3135	Rv3621c	manB
Rv0741	Rv1199c	Rv1790	Rv2490c	Rv3136	Rv3622c	ndh
Rv0742	Rv1214c	Rv1791	Rv2512c	Rv3144c	Rv3636	nat
Rv0746	Rv1243c	Rv1800	Rv2519	Rv3159c	Rv3637	oxyR
Rv0747	Rv1313c	Rv1801	Rv2591	Rv3184	Rv3638	pncA
Rv0754	Rv1325c	Rv1802	Rv2608	Rv3185	Rv3640c	rmlD
Rv0755A	Rv1361c	Rv1803c	Rv2615c	Rv3186	Rv3650	rpoB
Rv0755c	Rv1369c	Rv1806	Rv2634c	Rv3187	Rv3652	rpsL
Rv0795	Rv1370c	Rv1807	Rv2646	Rv3191c	Rv3653	rrs
Rv0796	Rv1386	Rv1808	Rv2648	Rv3325	Rv3738c	Rv0340
Rv0797	Rv1387	Rv1809	Rv2649	Rv3326	Rv3739c	Rv1592c
Rv0832	Rv1396c	Rv1818c	Rv2650c	Rv3327	Rv3746c	Rv1772
Rv0833	Rv1430	Rv1840c	Rv2651c	Rv3343c	Rv3751	Rv2242
Rv0834c	Rv1441c	Rv1917c	Rv2652c	Rv3344c	Rv3798	Rv3124
Rv0850	Rv1450c	Rv1918c	Rv2653c	Rv3345c	Rv3812	Rv3125c
Rv0872c	Rv1452c	Rv1983	Rv2654c	Rv3347c	Rv3827c	Rv3126c
Rv0878c	Rv1468c	Rv2013	Rv2655c	Rv3348	Rv3844	thyA
Rv0915c	Rv1548c	Rv2014	Rv2656c	Rv3349c	Rv3872	tlyA
Rv0916c	Rv1573	Rv2105	Rv2657c	Rv3350c	Rv3873	accD6
Rv0920c	Rv1574	Rv2106	Rv2659c	Rv3367	Rv3892c	
			Rv2666	Rv3380c	Rv3893c	

Listed are the genes that were excluded from the multiple alignment used to create the phylogenetic tree.

References

1. WHO. Global Tuberculosis Report 2015. World Health Organization; 2015.
2. Ganiem AR, Parwati I, Wisaksana R, van der Zanden A, van de Beek D, Sturm P, et al. The effect of HIV infection on adult meningitis in Indonesia: a prospective cohort study. AIDS. 2009;23(17):2309-16.
3. van Laarhoven AD, S.; Ruesen, C.; Hayati, E.; Damen, M. S. M. A.; Annisa, J.; Chaidir, L.; Netea, M. G.; Alisjahbana, B.; Ganiem, A. R.; van Crevel, R. Clinical parameters, routine inflammatory markers and LTA4H genotype as predictors for mortality among 608 tuberculous meningitis patients in Indonesia. The Journal of Infectious Diseases. 2017;In press.
4. Graustein AD, Horne DJ, Arentz M, Bang ND, Chau TT, Thwaites GE, et al. TLR9 gene region polymorphisms and susceptibility to tuberculosis in Vietnam. Tuberculosis (Edinb). 2015;95(2):190-6.
5. Campo M, Randhawa AK, Dunstan S, Farrar J, Caws M, Bang ND, et al. Common polymorphisms in the CD43 gene region are associated with tuberculosis disease and mortality. Am J Respir Cell Mol Biol. 2015; 52(3):342-8.
6. Hawn TR, Dunstan SJ, Thwaites GE, Simmons CP, Thuong NT, Lan NT, et al. A polymorphism in Toll-interleukin 1 receptor domain containing adaptor protein is associated with susceptibility to meningeal tuberculosis. J Infect Dis. 2006;194(8):1127-34.
7. Thuong NT, Hawn TR, Thwaites GE, Chau TT, Lan NT, Quy HT, et al. A polymorphism in human TLR2 is associated with increased susceptibility to tuberculous meningitis. Genes Immun. 2007;8(5):422-8.
8. Hoal-Van Helden EG, Epstein J, Victor TC, Hon D, Lewis LA, Beyers N, et al. Mannose-binding protein B allele confers protection against tuberculous meningitis. Pediatr Res. 1999;45(4 Pt 1):459-64.
9. Coscolla M, Gagneux S. Consequences of genomic diversity in *Mycobacterium tuberculosis*. Semin Immunol. 2014;26(6):431-44.
10. Black PA, de Vos M, Louw GE, van der Merwe RG, Dippenaar A, Streicher EM, et al. Whole genome sequencing reveals genomic heterogeneity and antibiotic purification in *Mycobacterium tuberculosis* isolates. BMC Genomics. 2015;16(1):857.
11. Guerra-Assuncao JA, Houben RM, Crampin AC, Mzembe T, Mallard K, Coll F, et al. Recurrence due to relapse or reinfection with *Mycobacterium tuberculosis*: a whole-genome sequencing approach in a large, population-based cohort with a high HIV infection prevalence and active follow-up. J Infect Dis. 2015;211(7):1154-63.
12. Reed MB, Domenech P, Manca C, Su H, Barczak AK, Kreiswirth BN, et al. A glycolipid of hypervirulent tuberculosis strains that inhibits the innate immune response. Nature. 2004;431(7004):84-7.
13. Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, Narayanan S, et al. Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. Proc Natl Acad Sci U S A. 2006;103(8):2869-73.
14. Guerra-Assuncao JA, Crampin AC, Houben RM, Mzembe T, Mallard K, Coll F, et al. Large-scale whole genome sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area. Elife. 2015;4.
15. de Jong BC, Hill PC, Aiken A, Awine T, Antonio M, Adetifa IM, et al. Progression to active tuberculosis, but not transmission, varies by *Mycobacterium tuberculosis* lineage in The Gambia. J Infect Dis. 2008;198(7): 1037-43.
16. van Crevel R, Nelwan RH, de Lenne W, Veeraragu Y, van der Zanden AG, Amin Z, et al. *Mycobacterium tuberculosis* Beijing genotype strains associated with febrile response to treatment. Emerg Infect Dis. 2001;7(5):880-3.
17. Parwati I, Alisjahbana B, Apriani L, Soetikno RD, Ottenhoff TH, van der Zanden AG, et al. *Mycobacterium tuberculosis* Beijing genotype is an independent risk factor for tuberculosis treatment failure in Indonesia. J Infect Dis. 2010;201(4):553-7.
18. Rakotosamimanana N, Raharimanga V, Andriamandimby SF, Soares JL, Doherty TM, Ratsitorahina M, et al. Variation in gamma interferon responses to different infecting strains of *Mycobacterium tuberculosis* in acid-fast bacillus smear-positive patients and household contacts in Antananarivo, Madagascar. Clin Vaccine Immunol. 2010;17(7):1094-103.
19. van Laarhoven A, Mandemakers JJ, Kleinnijenhuis J, Enaimi M, Lachmandas E, Joosten LA, et al. Low induction of proinflammatory cytokines parallels evolutionary success of modern strains within the *Mycobacterium tuberculosis* Beijing genotype. Infect Immun. 2013;81(10):3750-6.

20. Portevin D, Gagneux S, Comas I, Young D. Human macrophage responses to clinical isolates from the *Mycobacterium tuberculosis* complex discriminate between ancient and modern lineages. PLoS Pathog. 2011;7(3):e1001307.
21. Sarkar R, Lenders L, Wilkinson KA, Wilkinson RJ, Nicol MP. Modern lineages of *Mycobacterium tuberculosis* exhibit lineage-specific patterns of growth and cytokine induction in human monocyte-derived macrophages. PLoS One. 2012;7(8):e43170.
22. Be NA, Lamichhane G, Grosset J, Tyagi S, Cheng QJ, Kim KS, et al. Murine model to study the invasion and survival of *Mycobacterium tuberculosis* in the central nervous system. J Infect Dis. 2008;198(10):1520-8.
23. Be NA, Bishai WR, Jain SK. Role of Mycobacterium tuberculosis *pknD* in the pathogenesis of central nervous system tuberculosis. BMC Microbiol. 2012;12:7.
24. Skerry C, Pokkali S, Pinn M, Be NA, Harper J, Karakousis PC, et al. Vaccination with recombinant *Mycobacterium tuberculosis PknD* attenuates bacterial dissemination to the brain in guinea pigs. PLoS One. 2013;8(6):e66310.
25. Hernandez Pando R, Aguilar D, Cohen I, Guerrero M, Ribon W, Acosta P, et al. Specific bacterial genotypes of *Mycobacterium tuberculosis* cause extensive dissemination and brain infection in an experimental model. Tuberculosis (Edinb). 2010;90(4):268-77.
26. Tsenova L, Ellison E, Harbacheuski R, Moreira AL, Kurepina N, Reed MB, et al. Virulence of selected *Mycobacterium tuberculosis* clinical isolates in the rabbit model of meningitis is dependent on phenolic glycolipid produced by the bacilli. J Infect Dis. 2005;192(1):98-106.
27. Jain SK, Paul-Satyaseela M, Lamichhane G, Kim KS, Bishai WR. *Mycobacterium tuberculosis* invasion and traversal across an in vitro human blood-brain barrier as a pathogenic mechanism for central nervous system tuberculosis. J Infect Dis. 2006;193(9):1287-95.
28. Click ES, Moonan PK, Winston CA, Cowan LS, Oeltmann JE. Relationship between *Mycobacterium tuberculosis* phylogenetic lineage and clinical site of tuberculosis. Clin Infect Dis. 2012;54(2):211-9.
29. Pareek M, Evans J, Innes J, Smith G, Hingley-Wilson S, Lougheed KE, et al. Ethnicity and mycobacterial lineage as determinants of tuberculosis disease phenotype. Thorax. 2013;68(3):221-9.
30. Firdessa R, Berg S, Hailu E, Schelling E, Gumi B, Erenso G, et al. Mycobacterial lineages causing pulmonary and extrapulmonary tuberculosis, Ethiopia. Emerg Infect Dis. 2013;19(3):460-3.
31. Nicol MP, Sola C, February B, Rastogi N, Steyn L, Wilkinson RJ. Distribution of strain families of *Mycobacterium tuberculosis* causing pulmonary and extrapulmonary disease in hospitalized children in Cape Town, South Africa. J Clin Microbiol. 2005;43(11):5779-81.
32. Caws M, Thwaites G, Dunstan S, Hawn TR, Lan NT, Thuong NT, et al. The influence of host and bacterial genotype on the development of disseminated disease with *Mycobacterium tuberculosis*. PLoS Pathog. 2008;4(3):e1000034.
33. Saw SH, Tan JL, Chan XY, Chan KG, Ngeow YF. Chromosomal rearrangements and protein globularity changes in *Mycobacterium tuberculosis* isolates from cerebrospinal fluid. PeerJ. 2016;4:e2484.
34. Coll F, Preston M, Guerra-Assuncao JA, Hill-Cawthorn G, Harris D, Perdigao J, et al. PolyTB: a genomic variation map for *Mycobacterium tuberculosis*. Tuberculosis (Edinb). 2014;94(3):346-54.
35. Walker TM, Ip CLC, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. The Lancet Infectious Diseases. 2013;13(2):137-46.
36. Chen PE, Shapiro BJ. The advent of genome-wide association studies for bacteria. Curr Opin Microbiol. 2015;25:17-24.
37. Sassetti CM, Rubin EJ. Genetic requirements for mycobacterial survival during infection. Proc Natl Acad Sci U S A. 2003;100(22):12989-94.
38. Zheng J, Ren X, Wei C, Yang J, Hu Y, Liu L, et al. Analysis of the secretome and identification of novel constituents from culture filtrate of bacillus Calmette-Guerin using high-resolution mass spectrometry. Mol Cell Proteomics. 2013;12(8):2081-95.
39. Nogueira T, Rankin DJ, Touchon M, Taddei F, Brown SP, Rocha EP. Horizontal gene transfer of the secretome drives the evolution of bacterial cooperation and virulence. Curr Biol. 2009;19(20):1683-91.

40. Lees JA, Kremer PH, Manso AS, Croucher NJ, Ferwerda B, Seron MV, et al. Large scale genomic analysis shows no evidence for pathogen adaptation between the blood and cerebrospinal fluid niches during bacterial meningitis. Microb Genom. 2017;3(1):e000103.
41. Gagneux S. Host-pathogen coevolution in human tuberculosis. Philos Trans R Soc Lond B Biol Sci. 2012;367(1590):850-9.
42. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, et al. Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. Nat Genet. 2013;45(10):1183-9.
43. Nebenzahl-Guimaraes H, van Laarhoven A, Farhat MR, Koeken VA, Mandemakers JJ, Zomer A, et al. Transmissible *Mycobacterium tuberculosis* Strains Share Genetic Markers and Immune Phenotypes. Am J Respir Crit Care Med. 2016.
44. Gottesman S. Micros for microbes: non-coding regulatory RNAs in bacteria. Trends Genet. 2005;21(7):399-404.
45. Deatherage DE, Barrick JE. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. Methods Mol Biol. 2014;1151:165-88.
46. Coll F, McNerney R, Preston MD, Guerra-Assuncao JA, Warry A, Hill-Cawthorne G, et al. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Med. 2015;7(1):51.
47. Farhat MR, Shapiro BJ, Sheppard SK, Colijn C, Murray M. A phylogeny-based sampling strategy and power calculator informs genome-wide associations study design for microbial pathogens. Genome Med. 2014;6(11):101.
48. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59(3):307-21.
49. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, et al. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res. 2014;42(Database issue):D581-91.
50. Ashkenazy H, Penn O, Doron-Faigenboim A, Cohen O, Cannarozzi G, Zomer O, et al. FastML: a web server for probabilistic reconstruction of ancestral sequences. Nucleic Acids Res. 2012;40(Web Server issue):W580-4.
51. Nurk S, Bankevich A, Antipov D, Gurevich A, Korobeynikov A, Lapidus A, et al. Assembling Genomes and Mini-metagenomes from Highly Chimeric Reads. 2013;7821:158-70.
52. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068-9.
53. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25(7):1043-55.
54. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24-6.
55. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res. 2005;33(Web Server issue):W306-10.
56. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248-9.
57. Emanuelsson O, Nielsen H, Brunak S, von Heijne G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol. 2000;300(4):1005-16.
58. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 2001;305(3):567-80.
59. Ministry of Health Indonesia. Directorate General of Disease Control and Environmental Health. Petunjuk Teknis Manajemen TB Anak. 2013.

6

Mycobacterium tuberculosis Beijing lineage evades BCG protection against infection

Ayesha J. Verrall, Carolien Ruesen, Lidya Chaidir, Lika Apriani,
James E. Ussher, Arjan van Laarhoven, Rovina Ruslami, Martijn A. Huynen,
Mihai G. Netea, Jakko van Ingen, Katrina Sharples, Philip C. Hill,
Reinout van Crevel, Bachti Alisjahbana.

In preparation

Abstract

Background

Many studies have examined the host and environmental factors that influence tuberculosis transmission and secondary cases of disease, but the *M. tuberculosis* genetic determinants of transmission remain poorly understood.

Methods

Using a household-based case-contact study in Bandung, Indonesia, we determined whether the genotype of the *M. tuberculosis* strain that a contact is exposed to influences their risk of infection, measured by interferon gamma release assay conversion. Generalised estimating equation for Poisson regression was used to investigate whether the *M. tuberculosis* genotype influences the disease phenotype in the contact. We calculated a transmission index reflecting the isolate's ability to transmit to contacts, adjusted for index case characteristics and exposure risk of the contact.

Results

M. tuberculosis genotype of 414 index case isolates was linked to IGRA conversion of 1,201 household contacts. The Beijing genotype was associated with a 39% increased risk of uninfected contacts to acquire *M. tuberculosis* infection compared to non-Beijing genotype strains (relative risk 1.39; 95% confidence interval 1.00 – 1.93; p=0.048). This was not explained by a higher bacillary load or higher frequency of pulmonary cavities in the index cases. In addition, BCG vaccination showed no protection against *M. tuberculosis* infection in contacts exposed to a Beijing genotype strain (RR=1.02; 95% CI 0.56 – 1.85; p=0.9), whereas a strong protection was observed against infection with non-Beijing genotype strains (RR 0.40; 95% CI 0.27 - 0.61; p<0.001).

Discussion

Beijing genotype strains are associated with increased transmission and BCG vaccination showed less protection in contacts exposed to Beijing genotype strains. Evading innate immune clearance in the contact, rather than inducing a disease phenotype in the index case that favours transmission, may explain the increased transmissibility of Beijing genotype strains.

Introduction

Mycobacterium tuberculosis is the causative agent of tuberculosis (TB), the number one cause of death from an infectious disease caused by a single pathogen[1]. Mounting evidence from countries with high TB burdens points to ongoing transmission as the driving force maintaining TB incidence[2,3]. Despite coordinated efforts to control TB transmission, an estimated 10 million people developed TB disease in 2017[1] and 50 million people worldwide are predicted to become newly infected with *M. tuberculosis* every year[4]. Transmission of *M. tuberculosis* occurs through aerosol droplets that are produced by coughing pulmonary TB patients and inhaled by exposed contacts. Tuberculosis transmission is dependent on the prevalence and infectiousness of pulmonary TB cases, the number and susceptibility of TB case contacts, the frequency and proximity of interactions between TB cases and contacts, as well as biological features of *M. tuberculosis*[5]. To date, considerable emphasis has been placed on the role of host and environmental factors associated with transmission, but far less effort has been put in understanding pathogen factors involved in manipulating and evading host immunity[6]. *M. tuberculosis* consists of seven human-adapted phylogenetic lineages and is subdivided into genetic families defined by genetic traits (single nucleotide polymorphisms (SNPs) or deletions)[7]. The East-Asian lineage (Lineage 2) is widespread globally and the Beijing genotype family is its major component[8], which can be further subdivided into ancient and modern Beijing lineages[9]. Beijing strains have been associated with large outbreaks[10-13], suggesting increased transmissibility for this genotype family.

M. tuberculosis genotype-specific traits could increase the risk of transmission in two ways: first; by inducing a disease phenotype in the index case that favours transmission, or second; by evading host immune responses that eradicate *M. tuberculosis* before an adaptive immune response develops[14]. In the current household case contact study, in which contacts are linked to a known index case, we sought to understand the influence of host and pathogen factors on *M. tuberculosis* transmission. Specifically we sought to test whether increased Beijing lineage transmission was related to a transmission-favourable disease phenotype or to evasion of host responses in the contact.

Methods

Setting and design

The Innate Factors and Early Clearance of *M. tuberculosis* (INFECT) study was conducted in Bandung, Indonesia (estimated TB incidence: 395/100,000)[1]. The TANDEM project[15], on the relationship between TB and diabetes mellitus, recruited sputum smear positive TB patients over 15 years of age with an abnormal chest x-ray. Patients were eligible for the INFECT study if they shared a household with others and had received less than two weeks of TB treatment. Their household contacts[16] were eligible if they had lived with the index case for more than five hours a week and had no previous TB. Children under five years were ineligible and were referred to a primary care physician. The study was approved by the Health Research Ethics Committee Universitas Padjadjaran (14/UN6.C2.1.2/KEPK/PN/2014) and the Southern Health and Disability Ethics Committee New Zealand (13/STH/132). All participants gave written informed consent.

Study procedures

Patients with suspected TB disease were referred from primary care. Demographic characteristics, sputum smear grade[17], and *M. tuberculosis* culture results were recorded[18]. Their chest x-rays were read by a physician who identified cavities and the extent of abnormalities, as previously described[19,20]. They were treated for TB at no cost by the National TB Control Programme. Trained nurses recruited eligible contacts at home. They recorded demographic characteristics, smoking status, diabetes, history of human immunodeficiency virus (HIV) infection, TB symptoms, and sleep proximity to and hours spent with the index case the day before enrolment. Physical examination included assessment of the presence of a Bacillus Calmette-Guérin (BCG) vaccination scar, height and weight. Household contact *M. tuberculosis* infection status was assessed by QuantiFERON®-TB Gold In-Tube interferon gamma release assay (IGRA), interpreted according to the manufacturer's instructions[21]. Indeterminate tests were repeated, and the repeat value used. A random capillary blood glucose (RCBG) was performed, followed by venous glycosylated haemoglobin (HbA1c) for those with RCBG › 100 mg/dL (5.5 mmol/L). Contacts who were IGRA negative at baseline had a repeat IGRA at 14 weeks. Those with TB symptoms at baseline or follow up underwent chest x-ray and sputum examination.

M. tuberculosis isolates were suspended in liquid media and stored at -80°C until they were thawed and re-cultured for DNA extraction. *M. tuberculosis* DNA was extracted after subculturing of positive culture on Ogawa solid medium using the cetyl trimethylammonium bromide (CTAB) method or the UltraClean® Microbial DNA Isolation Kit (MO BIO Laboratories, Carlsbad, CA).

M. tuberculosis DNA from 137 Indonesian isolates was sequenced on an Illumina HiSeq 2000 instrument using 2 x 100 bp paired-end reads at the Beijing Genome Institute in Hong Kong; the remaining 285 isolates were sequenced on an Illumina NextSeq 500 instrument using 2 x 150 bp paired-end reads at the department of Human Genetics at the Radboudumc, Nijmegen, the Netherlands. After sequencing, the raw FASTQ sequence reads were filtered, including removing of adapter sequences. All sequencing reads passed the Illumina base caller filter. Sequencing coverage was determined using the FastQC quality control tool version 0.10.1, and the Genome Analysis Toolkit[22]. Sequencing coverage statistics are shown in supplementary File 1. The sequence reads were aligned to reference strain *M. tuberculosis* H37Rv, accession number NC_000962.3, and variants were called using Breseq, version 0.33.2[23]. We extracted all 57,045 variable positions across the 414 *M. tuberculosis* sequences and concatenated them into a multiple sequence alignment. Solely for the purpose of creating the phylogenetic tree, SNPs occurring in PE/PPE genes, genes related to mobile elements, as well as genes previously associated with drug resistance[24] were removed. The remaining 55,494 SNPs were used to construct the phylogenetic tree using PhyML version 3.3.20180621[25] using the HKY85 model with four gamma-distributed rate-categories, and using a hundred bootstraps. We used the R package ape[26] (version 5.3) to calculate the distance to the genetically closest other isolate (minimum pairwise distance) for all isolates, again excluding SNPs in PE/PPE genes, genes related to mobile elements and genes associated with drug resistance. We determined *M. tuberculosis* lineage based on a 62-single nucleotide polymorphisms (SNPs) barcode[27]. To further classify isolates belonging to lineage 2 (East-Asian lineage) into non-Beijing, ancient Beijing and modern Beijing sublineages, we categorized isolates belonging to lineage 2.1 as non-Beijing (or proto-Beijing), those belonging to lineage 2.2.2 as ancient Beijing, and those belonging to lineage 2.2.1.1 or 2.2.1.2 as modern Beijing. The lineage 2.2.1 isolates, which cannot be unambiguously distinguished into ancient or modern Beijing according to the 62-SNP-barcode, were categorized according to the *M. tuberculosis* lineage 2 dendrogram published by Shitikov *et al*[28]. We used TBProfiler version 2.1.1[24] to determine genotypic drug resistance. This tool uses raw sequence data as input, aligns these to the *M. tuberculosis* H37Rv reference genome, and compares identified SNPs and indels to a curated list of drug resistance mutations.

Data management and analysis

All data were double entered into a database (Microsoft Access) and checked for errors. IGRA conversion was defined as a change from a negative baseline test to positive at 14-week follow-up. Body mass index (BMI) was calculated according to the formula BMI = weight/(height^2). Diabetes was defined as an HbA1c^{3}6.5% and pre-diabetes as HbA1c 5.7-6.4%. To create one summary measure of exposure to

facilitate specific analyses described below, exposure risk scores were predicted from a logistic regression of *M. tuberculosis* exposure variables (index case: sputum smear grade, drug resistance mutations, cavities, extent of x-ray disease; contacts: hours spent with, and sleeping proximity to, the case) against IGRA results, as described previously[29].

The primary analysis was of risk of IGRA conversion for those with a negative IGRA at baseline, as incident conversions are almost certainly due to exposure to the index case. A sensitivity analysis of risk of a positive IGRA at baseline or follow up, compared to a persistently negative IGRA, was also performed.

Associations between host factors and IGRA were estimated using Poisson regression generalised estimating equations, clustering on case, and with robust standard errors to account for binomial data. Index case characteristics and exposure risk score were included as potential confounders. A modified backwards stepwise regression was used, with age and sex retained, and other variables retained if $p<0.2$. Estimates for baseline associations are presented as prevalence ratios (PR) and those for conversion are presented as relative risks (RR).

To study pathogen genetic determinates of BCG evasion we calculated a transmission index that reflects the isolate's ability to transmit to contacts, accounting for the factors extrinsic to the pathogen such as contact susceptibility and proximity to the index case. The transmission index was calculated by subtracting expected instances of transmission to BCG-vaccinated contacts from the observed instances of transmission to BCG-vaccinated contacts in each household. To score observed transmission events we weighted a positive IGRA at baseline at 20% of a conversion, as there is less certainty that a positive IGRA at baseline resulted from exposure to the index case. The expected risk of transmission was predicted using the multivariable model for the primary analysis, but with the lineage term removed. All analyses were carried out using Stata version 15.

Results

TB household index cases and contacts

From home visits to 465 index case households, 2,090 contacts were screened and 1,620 were eligible (Figure 6.1). Of these, 1,347 (83%) contacts of 462 index cases gave informed consent for inclusion in the study. The whole genome sequencing was available on 414 index case isolates linked to 1,201 contacts. The remaining 48 index cases were not included because they did not have a culture submitted (n=13),

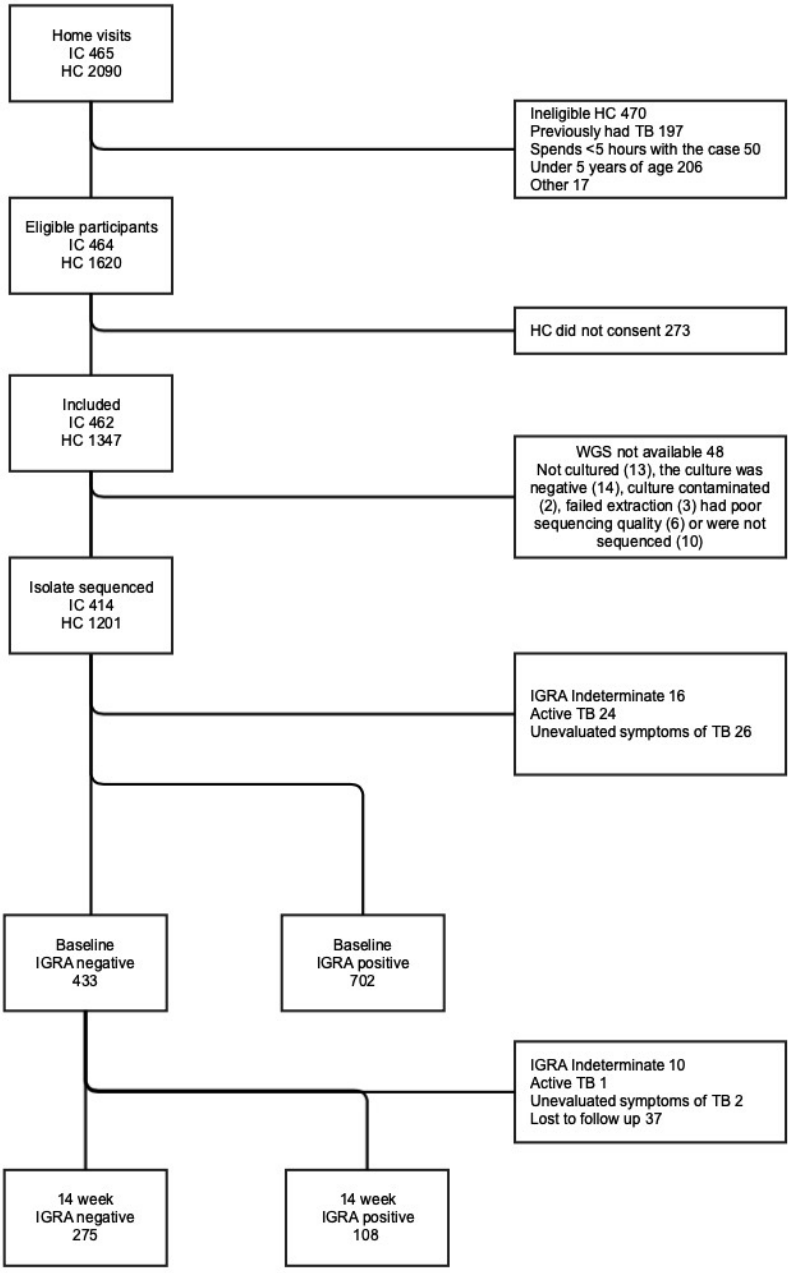


Figure 6.1. Recruitment and enrolment of participants into the study and outcome of participants included in the study.

the culture was negative (n=14), or contaminated (n=2), failed extraction (n=3) had poor sequencing quality (n=6) or were not sequenced (n=10). Contacts were classified into three groups as: baseline IGRA positive (n=709), persistently IGRA negative (n=277), and IGRA converters (n=108). For the 414 isolates that were whole genome sequenced, the median coverage depth was 96x, (range 15-363) and the median percentage of bases covered by at least one read was 99.2%.

The three classified groups of case contacts were similar with respect to median age, sex and duration of treatment of their respective index cases and with respect to their sex, smoking history and diabetes measurements (Table 6.1). However, persistently IGRA negative contacts had relatively lower measures of exposure to *M. tuberculosi*s and, along with converters, were younger than those who were initially IGRA positive.

Of the sequenced index case isolates, 260 (62.8%) belonged to the Euro-American lineage, 133 (32.1%) were East-Asian (Beijing), seven (1.7%) were East-Asian lineage but not from the Beijing lineage and 14 (3.4%) were from the Indo-Oceanic lineage (Figure 6.2). Fifty-five (41%) of the East-Asian Beijing genotype isolates were ancient Beijing isolates, and 78 (59%) were modern Beijing isolates. Two (Euro-American) isolates differed by only ten SNPs, and could have been part of the same transmission cluster[30].

The clinical or radiological characteristics of index cases were similar for those infected with Beijing isolates compared to those infected with non-Beijing isolates (Table 6.2). Beijing isolates were more likely to be multi-drug resistant than non-Beijing isolates (5.3 vs. 1.8% respectively, P=0.046). Table S6.1 summarizes drug resistance mutations according to lineage. Patients infected with Beijing and non-Beijing isolates were similar with respect to the smear grade and presence of cavities. There was no difference in the dominant abnormality in each lung zone, according to lineage (Table S6.2).

Table 6.1. Characteristics of 1,085 contacts and their index case according to the contact's IGRA result

Characteristic	Baseline IGRA Positive	Baseline IGRA Negative	
	(N=702)	IGRA converter (N=108)	IGRA Persistently Negative (N=275)
Contact's index case[a]			
Age (years) median (IQR)	37.6 (28.6 - 49.8)	38.4 (28.7 - 50.8)	40.9 (30.5 - 51.9)
Female	323 (46.0)	56 (51.9)	144 (52.4)
Days of treatment			
None	574 (81.8)	87 (80.6)	224 (81.5)
<7	115 (16.4)	20 (18.5)	46 (16.7)
7 to 13	13 (1.9)	1 (0.9)	5 (1.8)
Diabetes[b]	169 (28.8)	36 (37.9)	84 (33.5)
Current smoker[b]	66 (11.2)	11 (11.6)	31 (12.4)
HIV[c]	1 (0.2)	2 (2.2)	0 (0.0)
Temperature[b] (C), median (IQR)	36.8 (36.4 - 37.3)	36.8 (36.4 - 37.3)	36.8 (36.4 - 37.4)
X-ray[d] cavities (present)	404 (59.5)	55 (51.9)	121 (46.7)
Extent of chest x-ray disease, median (IQR)	50.0 (25.0 - 75.0)	40.0 (25.0 - 55.0)	40.0 (25.0 - 60.0)
Sputum smear grade			
3+	345 (49.1)	60 (55.6)	90 (32.7)
2+	199 (28.3)	26 (24.1)	76 (27.6)
Scanty/1+	158 (22.5)	22 (20.4)	109 (39.6)
Any drug resistance	77 (11.0)	17 (15.7)	30 (10.9)
MDR	17 (2.4)	2 (1.9)	10 (3.6)
Beijing lineage	246 (35.0)	39 (36.0)	69 (25.0)
Contacts			
Hours with the case median (IQR)	5.0 (2.0 - 10.0)	5.0 (3.0 - 8.0)	3.0 (1.0 - 7.0)
Sleep proximity to index case			
Same room	482 (68.7)	88 (81.5)	216 (78.5)
Different room	220 (31.3)	20 (18.5)	59 (21.5)
Age (years) median (IQR)	30 (16 - 46)	23 (15 - 35)	23 (13 - 41)
Female	408 (58.1)	57 (52.8)	149 (54.2)
BCG vaccination	539 (76.8)	80 (74.1)	240 (87.3)

Table 6.1. Continued.

Characteristic	Baseline IGRA Positive	Baseline IGRA Negative	
	(N=702)	IGRA converter (N=108)	IGRA Persistently Negative (N=275)
Smoking history			
Never smoked	458 (65.2)	70 (65.4)	194 (70.5)
Quit >6 months ago	30 (4.3)	3 (2.8)	13 (4.7)
Current smoker	214 (30.5)	34 (31.8)	68 (24.7)
BMI (kg/m^2), mean (SD)	21.9 (5.3)	21.6 (5.3)	21.0 (5.0)
Diabetes[e]			
No diabetes	637 (90.7)	96 (88.9)	253 (92.0)
Pre-diabetes	44 (6.3)	7 (6.5)	13 (4.7)
Diabetes	21 (3.0)	5 (4.6)	9 (3.3)

Figures are n (%) unless otherwise indicated.
[a]Index cases may be represented multiple times if they have more than one contact.
[b]index case smoking, diabetes and temperature available for index cases of 1,036 contacts.
[c]HIV result available for index cases of 1,025 contacts.
[d]X-ray available for detailed reading in index cases of 1,158 contacts.
[e]Diabetes defined as follows: No diabetes: RCBG<101 or HbA1c<5.7%; Pre-diabetes: HbA1c 5.7%-6.4%; Diabetes: HbA1c≥6.5%
Abbreviations: PN: persistently negative; IGRA: interferon-gamma release assay; BMI: body mass index; HIV: human immunodeficiency virus; BCG: Bacillus Calmette–Guérin. SD: standard deviation; IQR: interquartile range; IGRA: interferon-gamma release assay

Index isolate and case contact IGRA results at follow-up

In a multivariable model, the relative risk (RR) of IGRA conversion was 1.39 (95% CI 1.00 – 1.93; p=0.048) for contacts exposed to Beijing versus non-Beijing isolates (Table 6.3). Risk of conversion was also increased in contacts exposed to high smear grade index cases and those who spent longer with the case, and was decreased in those who were BCG vaccinated. The effect of BCG vaccination was modified by the lineage of the isolate the contact was exposed to ($p_{interaction}$=0.01). BCG showed strong protection for contacts exposed to non-Beijing lineages (RR 0.40; 95% CI 0.27 - 0.61; p<0.001), whereas no protection was observed for contacts exposed to Beijing lineage (RR=1.02; 95% CI 0.56 – 1.85; p=0.9) (Table 6.4).

Table 6.2. Clinical and x-ray characteristics of 414 index cases according to lineage.

Characteristic	Non-Beijing (N=282)	Beijing (N=132)	Total	P-value
Age (years)	40.7 (14.6)	38.2 (12.9)	39.9 (14.1)	0.09
Female sex	139 (49.3)	55 (41.7)	194 (46.9)	0.1
Cough	277 (98.2)	131 (99.2)	408 (98.6)	0.4
HIV	0 (0.0)	1 (0.9)	1 (0.3)	0.1
Diabetes	68 (27.8)	34 (30.6)	102 (28.7)	0.5
Current smoker	28 (11.4)	10 (9.0)	38 (10.7)	0.5
Temperature	36.9 (0.8)	36.8 (0.7)	36.9 (0.8)	0.3
Sputum smear grade				0.793
3+	85 (30.1)	38 (28.8)	123 (29.7)	
2+	76 (27.0)	36 (27.3)	112 (27.1)	
Scanty/1+	121 (42.9)	58 (43.9)	179 (43.2)	
Any drug resistance	40 (14.2)	19 (14.4)	59 (14.3)	0.9
MDR	5 (1.8)	7 (5.3)	12 (2.9)	0.046
Cavities	142 (52.4)	70 (53.8)	212 (52.9)	0.8
Temperature	45.0 (25.0-70)	40.0 (25.0-65)	45.0 (25.0-70)	0.3
Hilar adenopathy	50 (18.5)	25 (19.2)	75 (18.7)	0.9
Aorto pulmonary window	42 (15.5)	24 (18.5)	66 (16.5)	0.5
Pleural effusion	23 (8.5)	7 (5.4)	30 (7.5)	0.3
Pleural thickening	14 (5.2)	7 (5.4)	21 (5.2)	0.9
Pleural calcification	3 (1.1)	0 (0.0)	3 (0.7)	0.2
Cardiomegaly	2 (0.7)	0 (0.0)	2 (0.5)	0.3
Tracheal deviation	28 (10.3)	18 (13.8)	46 (11.5)	0.3
Right upper lobe collapse	53 (19.6)	30 (23.1)	83 (20.7)	0.4
Left upper lobe collapse	26 (9.6)	12 (9.2)	38 (9.5)	0.9
Miliary pattern	3 (1.1)	2 (1.5)	5 (1.2)	0.7
Calcified granuloma	233 (86.0)	116 (89.2)	349 (87.0)	0.4
Multiple calcified granuloma	20 (7.4)	8 (6.2)	28 (7.0)	0.7

X-ray available on 401 of 414 index cases. Smoking, diabetes and temperature available for 356 index cases. HIV result available for 351 index cases.
Abbreviations: MDR, multidrug-resistant; HIV, human immunodeficiency virus

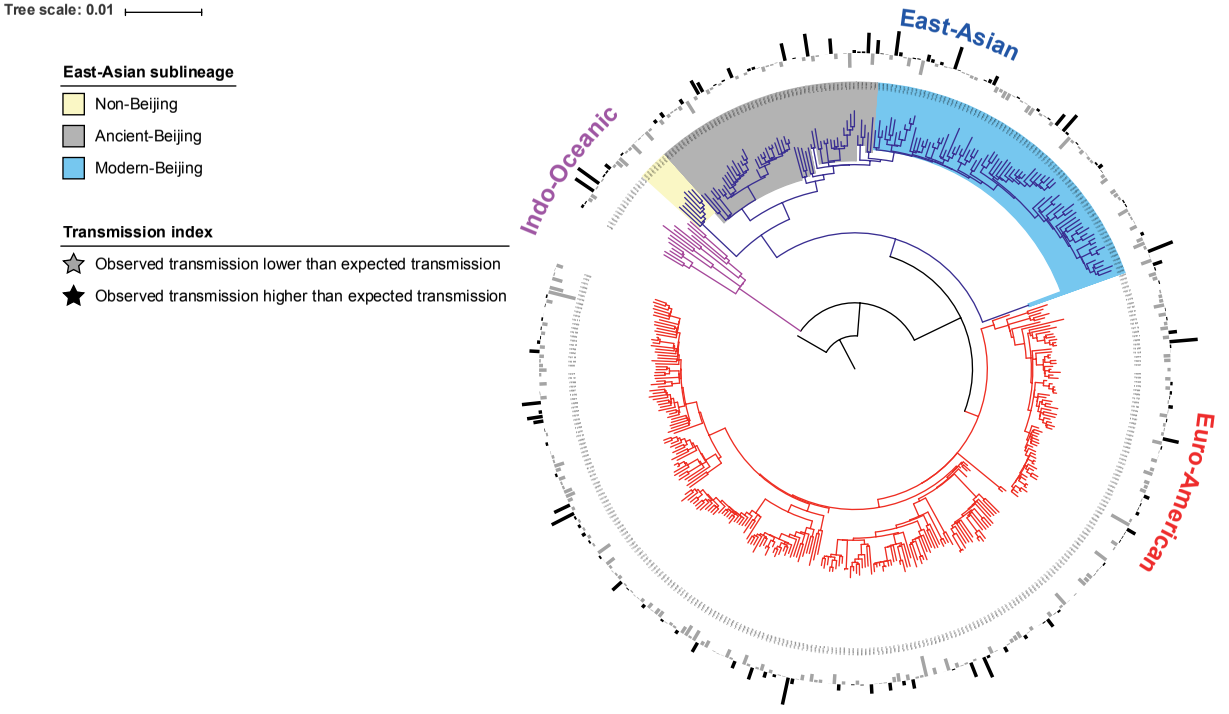


Figure 6.2. Phylogenetic tree of 414 successfully sequenced *M. tuberculosis* isolates with transmission indexes for the 373 isolates with BCG-vaccinated contacts . Bars in the outer ring indicate the transmission indexes.

Sensitivity analysis

The risk of a positive IGRA at either baseline or follow up was increased for contacts of index cases with a Beijing lineage isolate and decreased for BCG-vaccinated contacts (Table S6.3). In a multivariable model the effect of BCG vaccination was modified by the lineage of the index cases' isolate ($p_{interaction}$=0.03). Contacts of index cases with a non-Beijing lineage had a reduced risk of a positive IGRA if BCG-vaccinated (RR 0.82; 95% CI 0.76 - 0.89; p<0.001), whereas BCG vaccination was not associated with protection for contacts of index cases with a Beijing lineage isolate (RR=0.96; 95% CI 0.86 – 1.06; p=0.4) (Table S6.4).

Exploratory analysis

Of contacts with a negative IGRA at baseline, 25 (9.1%) were exposed to an ancient Beijing isolate, 44 (16.0%) to a modern Beijing isolate and 206 (74.9%) to non-Beijing isolates. Of contacts who converted, 19 (17.6%) were exposed to an ancient Beijing isolate, 20 (18.5%) to a modern Beijing isolate, and 69 (63.9%) to non-Beijing isolates. At follow up, the risk of IGRA conversion was 1.72 (95% CI 1.02 – 2.89; P=0.041) for contacts exposed to an ancient Beijing isolate versus non-Beijing and 1.25 (95% 0.77-2.01; P=0.370) for contacts exposed to a modern Beijing isolate versus non-Beijing (Table S6.5). Small numbers did not permit analysis of interaction with BCG.

Analysis of *M. tuberculosis* transmissibility

For the calculation of the *M. tuberculosis* transmission index, the total number of included households was 373; we excluded 41 without BCG-vaccinated contacts. We calculated an expected risk of conversion for 880 BCG-vaccinated contacts (BP: n=539; converter: n=80; PN: n=240; active TB: n=21). The mean transmission score for Beijing isolates was slightly higher but not significantly different from the mean transmission score for non-Beijing isolates (0.0037 vs. -0.0323, respectively; p=0.251, data not shown). Figure 6.2 shows the transmission indexes along the phylogenetic tree.

Table 6.3. Assessment of risk factors for IGRA conversion in household contacts who were IGRA negative at baseline (n=383)

	IGRA PN (N=275)	IGRA positive (N=108)	Crude RR	95% CI	P-value	ARR[a]	95% CI	P-value
Age index[e]	40.9 (30.3 - 51.9)	38.4 (28.7 - 50.8)	1.00	0.98 - 1.01	0.5	1.00	0.98 - 1.01	0.4
Index case sex								
Female	144 (52.0)	56 (51.9)	1.00	ref				
Male	133 (48.0)	52 (48.1)	1.01	0.71 - 1.45	0.9	0.97	0.69 -1.34	0.8
Diabetes index [b]								
No	167 (66.5)	59 (62.1)	1.00	ref				
Yes	84 (33.5)	36 (37.9)	1.15	0.77 - 1.71	0.491			
Smoking index [b]								
Not currently	220 (87.6)	84 (88.4)	1.00	ref				
Currently	31 (12.4)	11 (11.6)	0.95	0.52 - 1.72	0.9			
Cough duration (days)[e]	30.0 (21.0 - 90.0)	30.0 (30.0 - 90.0)	1.00	1.00 - 1.00	0.9			
Temperature[b] (°C) mean (SD)	36.9 (0.8)	36.8 (0.7)	0.90	0.71 - 1.15	0.4			
M. tuberculosis lineage								
Non-Beijing	208 (75.1)	69 (63.9)	1.00	ref				
Beijing	69 (24.9)	39 (36.1)	1.44	0.98 - 2.10	0.06	1.39	1.00 - 1.93	0.048
Index case highest smear grade								
Scanty1+	109 (39.4)	22 (20.4)	1.00	ref				
2+	76 (27.4)	26 (24.1)	1.52	0.89 - 2.59	0.1	1.28	0.76 - 2.15	0.3
3+	92 (33.2)	60 (55.6)	2.38	1.48 - 3.83	<0.05	2.26	1.43 - 3.59	0.001
Any drug resistance								
No	245 (89.1)	91 (84.3)	1.00	ref				
Yes	30 (10.9)	17 (15.7)	1.34	0.83 - 2.14	0.2			
Multidrug resistance								
No	265 (96.4)	106 (98.1)	1.00	ref				
Yes	10 (3.6)	2 (1.9)	0.58	0.19 - 1.75	0.3			
Cavities[c]								
No cavity	138 (52.9)	51 (48.1)	1.00	ref				
Present	123 (47.1)	55 (51.9)	1.16	0.82 - 1.64	0.4			

Extent x-ray disease[†ce]	40.0 (25.0 - 60.0)	40.0 (25.0 - 55.0)	1.00	0.99 - 1.01	0.8			
Age contact[e]	23.3 (12.6 - 41.5)	22.8 (14.6 - 35.5)	0.96	0.87 - 1.06	0.5	0.89	0.79 - 1.00	0.06
Contact sex								
Female	150 (54.2)	57 (52.8)	1.00	ref				
Male	127 (45.8)	51 (47.2)	1.04	0.75 - 1.45	0.8	0.99	0.68 - 1.43	0.9
Contact ethnicity								
Sunda	254 (91.7)	97 (89.8)	1.00	ref				
Other	23 (8.3)	11 (10.2)	1.03	0.95 - 1.11	0.5			
Sleep proximity to index								
Different Room	218 (78.7)	88 (81.5)	1.00	ref				
Same Room	59 (21.3)	20 (18.5)	0.87	0.57 - 1.34	0.5			
Hours with Case[e]	3.0 (1.0 - 7.0)	5.0 (3.0 - 8.0)	1.06	1.02 - 1.10	0.004	1.08	1.03 - 1.12	<0.001
BCG vaccination								
No	35 (12.6)	28 (25.9)	1.00	ref				
Yes	242 (87.4)	80 (74.1)	0.56	0.40 - 0.80	0.001	0.55	0.39 - 0.77	0.001
Smoking								
Non smoker	208 (75.1)	73 (68.2)	1.00	ref				
Current smoker	69 (24.9)	34 (31.8)	1.28	0.92, 1.78	0.1	1.43	0.97,2.09	0.07
Diabetes[d]								
No diabetes	254 (91.7)	96 (88.9)	1.00	ref				
Pre-diabetes	14 (5.1)	7 (6.5)	1.27	0.68, 2.40	0.456			
Diabetes	9 (3.2)	5 (4.6)	1.30	0.63, 2.67	0.479			
BMI[e]	20.2 (17.0 - 24.3)	21.0 (17.4 - 24.5)	1.02	0.98, 1.05	0.337	1.03	1.00,1.07	0.059

All figures are N (%) unless otherwise stated. Index cases maybe represented multiple times if they have more than one contact.
[a]Estimates obtained by multiple regression, adjusted for index case age, sex and smear grade, and contact age, sex, hours spent with index case smoking and body mass index.
[b]index case smoking, diabetes and temperature available for index cases of 346 contacts.
[c]X-ray available for detailed reading in index cases of 365 contacts.
[d]Diabetes defined as follows: No diabetes: RCBG<101 or HbA1c<5.7%; Pre-diabetes: HbA1c 5.7%-6.4%; Diabetes: HbA1c³6.5%.
[e]Median (interquartile range).
Abbreviations: PN: persistently negative; IGRA: interferon-gamma release assay; RR: relative risk; ARR: adjusted relative risk; 95% CI: 95% confidence interval; BCG: Bacillus Calmette-Guérin.

Table 6.4. Association between BCG vaccination and IGRA at 14 weeks for those negative at baseline, according to index case lineage

Lineage	BCG vaccination	IGRA PN N=275	IGRA positive N=108	RR[a]	95% CI	P-value
Non-Beijing	No	22	21	1.00	ref	
	Yes	184	48	0.40	0.27 - 0.61	<0.001
Beijing	No	13	7	1.00	ref	
	Yes	56	32	1.02	0.56 - 1.85	0.946

*Data are presented as N (%) unless indicated otherwise.
[a]Estimates obtained by multiple regression, adjusted for index case age, sex and smear grade, and contact age, sex, hours spent with index case smoking and body mass index.
Abbreviations: PN: persistently negative; IGRA: interferon-gamma release assay; BCG: Bacillus Calmette–Guérin; IGRA: interferon-gamma release assay; RR: relative risk calculated as risk ratios; 95% CI: 95% confidence interval.

Discussion

Here we have shown that both pathogen and host factors affect *M. tuberculosis* transmission. TB household contacts exposed to index cases infected with an *M. tuberculosis* Beijing isolate had a higher risk of becoming infected compared to those exposed to non-Beijing isolates. In addition, a history of BCG vaccination among TB household contacts was associated with a 60% lower risk of *M. tuberculosis* infection caused by non-Beijing isolates, but not with a lower risk of infection by Beijing family isolates. This relationship was independent of the presence of drug resistance mutations, or index case cavities or sputum smear grade. Therefore transmission of Beijing family isolates most likely arises from its ability to evade BCG-mediated protection in contacts, and not by inducing a more infectious disease phenotype in the index case[31] or through its association with drug resistance.

People have speculated that BCG vaccination beginning in the mid 20th century drove the evolution of the virulent modern Beijing lineage[32,33]. We observed that the isolates' transmissibility varied considerably along the phylogeny and within the Beijing lineage, and we did not find evidence of an evolutionary event after which the isolates became more transmissible (Figure 6.2). Rather, transmissibility seems to be a convergent trait, which has evolved independently multiple times. In addition, we did not find evidence to suggest that evasion of BCG protection was specific to modern Beijing family isolates. In fact, some studies have suggested that the evolution of

modern lineages predates the widespread use of BCG. Still, evasion of BCG-mediated immunity could contribute to the global emergence of the Beijing lineage. It may also contribute to the variability in estimates for protective effectiveness of BCG in different locations.

Our finding of an increased prevalence of *M. tuberculosis* infection in Indonesian case contacts of index cases with Beijing lineage isolates compared to others, mostly Euro-American lineage isolates, is consistent with the global emergence of the Beijing lineage over recent decades[10-13,34]. However, the results of the few case contact studies that have assessed the impact of the pathogen on transmission are not consistent. In a cohort study of patients with TB and their household contacts in The Gambia, no difference in transmission between the MTBC lineages was observed[35]; in addition, the ratio of infected child contacts among those exposed to a Beijing strain did not differ from those exposed to a non-Beijing strain in South-Africa[36]; and the Beijing genotype was not associated with a higher number of tuberculin skin test (TST)-positive contacts in The Netherlands[37]. In these countries however, the prevalence of the Beijing lineage is low (3 to 5% in Central America, Europe, and Africa[38]) and the estimates for the Beijing lineage may lack precision. A recent study of households in Guangxi, China, where the prevalence of Beijing genotype strains is higher (as high as 45% in Far East Asia[38]), showed that exposure to a Beijing strain was independently associated with a positive TST[39]. The discrepancies between results from these studies might also result from differences in how and if host-related factors were controlled for.

Another method that has been used to study transmission is investigating transmission chains, usually by clustering strains based on genotype similarities. Several studies have found increased clustering of Beijing lineage isolates, assessed by whole genome sequencing[10,13,40] or VNTR genotyping[41]. In contrast, two studies performed in low prevalence countries have not found increased clustering associated with the Beijing lineage[42,43]. The Beijing genotype consists of a number of sub-lineages and so the discrepant findings could be due to the prevalence of different sub-lineages in the different study populations, as suggested previously[44]. Within the Beijing lineage, modern Beijing strains were more often in a cluster than ancient Beijing strains in a study of 376 MDR-*M. tuberculosis* strains in China[45], and a study conducted in South Africa found significantly more clustering linked to recently evolved sub-lineages of the Beijing strain family compared to other sub-lineages[46]. These studies have inferred transmission from the frequency of clustered isolates in surveillance culture collections where clinical and exposure data on cases are often limited. Furthermore, clustering in such studies may arise from more efficient *transmission* of a particular isolate but also from differences in *progression* of latent TB infection to active disease. A strength of

the present study is the direct measurement of *transmission*, defined as passing infection from index case to contact, and the ability to find detailed data on the index, contacts and their exposure to one another, at a level of detail that is not feasible for studies of clustered isolates. The prospectively defined groups enabled better measurement of confounding factors and accounted for misclassification of baseline positive patients, whose source of infection might have been different than the household index case[47], and whose infection might have happened in the distant past.

A study has found that BCG vaccination and BMI were risk factors for clustering among Beijing strains in China[48]; both factors associated with a persistently negative IGRA in the present study. Adaptation of *M. tuberculosis* Beijing strains to their East-Asian host population as a result of host-pathogen co-evolution could have explained their higher transmissibility[49-51], but is unlikely to have caused the differential protective effect of BCG. It is difficult to speculate on the mechanism through which Beijing lineage strains evade BCG-mediated protection, when the mechanism of BCG protection itself is unknown. Beijing lineage strains could evade host immune response in the contact by evading recognition and reducing the pro-inflammatory cytokine production during infection, as seen in *ex vivo* studies[52-54]. In this cohort we previously showed that a persistently negative IGRA was associated with heterologous cytokine responses, similar to those activated during induction of trained innate immunity[55].

Without a safe human challenge model for tuberculosis, variation in BCG efficacy according to *M. tuberculosis* lineage cannot be assessed in experimental studies. Results from observational studies such as this need to be considered in light of their potential limitations. We determined BCG vaccination status by assessing the presence of a scar, as it is the most reliable method when dealing with a mixed age cohort. A minority of BCG vaccine recipients do not form a scar when administered at birth. Although this may introduce a bias whereby vaccination status inferred from scar formation leads to an underestimate of the effectiveness of BCG[56], it is not plausible that this misclassification differs according to the *M. tuberculosis* lineage of the index case, which was not determined until after the study concluded. Similarly, the IGRA is not a perfect test for *M. tuberculosis* infection, but there is no evidence that its performance differs according to infecting *M. tuberculosis* genotype. We have previously reported that more stringent definitions of conversions did not alter estimates of BCG mediated protection[29]. A sensitivity analysis that included participants with a positive IGRA at baseline replicated the main finding, indicating that the exclusion of these participants from the primary analysis was not an important source of bias. Finally, the distribution of lineages and size of the study meant only an analysis of Beijing lineage versus all other lineages was possible.

In summary, *M. tuberculosis* Beijing genotype strains are associated with more transmission and may evade BCG-induced immunity. Whether this finding is generalizable to other ethnic host populations warrants further investigation. The impact of BCG vaccination on TB transmission may depend on the *M. tuberculosis* lineage distribution in the setting it is used. Ultimately, understanding the underlying mechanism of BCG evasion by Beijing family isolates could advance the knowledge on the development of new vaccines.

Supplementary tables

Table S6.1. Drug resistance mutations summarized per gene in Beijing and non-Beijing *M. tuberculosis* isolates

Drug	Gene containing mutation(s)	Beijing (N=133)	Non-Beijing (N=281)
Isoniazid	*katG*	10 (8%)	14 (5%)
	Rv1482c-fabG1	2 (2%)	13 (5%)
	ahpC	1 (1%)	0
Rifampicin	*rpoB*	9 (7%)	6 (2%)
Ethambutol	*embB*	5 (4%)	3 (1%)
	embC-embA	0	1 (0.4%)
Pyrazinamide	*pncA*	4 (3%)	13 (5%)
Streptomycin	*rpsL*	5 (4%)	1 (0.4%)
	rrs	2 (2%)	0
	gid	0	2 (1%)
Ethionamide	*Rv1482c-fabG1*	2 (2%)	13 (5%)
	ethA	0	2 (1%)
Fluoroquinolones	*gyrA*	2 (2%)	1 (0.4%)
Amikacin	*rrs*	1 (1%)	0
Capreomycin	*rrs*	1 (1%)	0
	tlyA	0	2 (1%)
Kanamycin	*eis-Rv2417c*	1 (1%)	0
	rrs	1 (1%)	0
Para-aminosalicylic acid	*folC*	1 (1%)	1 (0.4%)

Data represent the number (%) of isolates with at least one known drug resistance-conferring mutation in the respective gene.

Table S6.2. Dominant x-ray abnormality according to lung zone and lineage for 401 index cases

	Non-Beijing				Beijing			
Dominant x-ray abnormality	**Any N (%)**	**Upper N (%)**	**Mid N (%)**	**Lower N (%)**	**Any N (%)**	**Upper N (%)**	**Mid N (%)**	**Lower N (%)**
None	2 (0.7)	21 (7.7)	32 (11.8)	89 (32.8)	2 (1.5)	19 (14.6)	26 (20.0)	53 (40.8)
Cavity	142 (52.4)	112 (41.3)	49 (18.1)	6 (2.2)	70 (53.8)	60 (46.2)	18 (13.8)	2 (1.5)
Consolidation	51 (18.8)	42 (15.5)	42 (15.5)	35 (12.9)	25 (19.2)	18 (13.8)	27 (20.8)	19 (14.6)
Patchy Lesion	74 (27.3)	89 (32.8)	137 (50.6)	134 (49.4)	32 (24.6)	32 (24.6)	56 (43.1)	52 (40.0)
Nodules	1 (0.4)	1 (0.4)	2 (0.7)	1 (0.4)	1 (0.8)	1 (0.8)	2 (1.5)	2 (1.5)
Bullae	1 (0.4)	6 (2.2)	9 (3.3)	6 (2.2)	0 (0.0)	0 (0.0)	1 (0.8)	2 (1.5)

Table S6.3. Assessment of risk factors for positive IGRA at baseline or follow-up (N=1,112)

	IGRA PN (N=276)	Any IGRA positive (N=836)	RR	95% CI	P-value	APR[a]	95% CI	P-value
Age index[b]	40.9 (30.5 - 51.9)	37.6 (28.6 - 49.8)	1.00	1.00 - 1.00	0.267	1.00	1.00 - 1.00	0.302
Index case sex								
Female	145 (52.5)	392 (46.9)	1.00	ref				
Male	131 (47.5)	444 (53.1)	1.06	0.96 - 1.16	0.232	1.03	0.95 - 1.12	0.43
Diabetes index								
No	167 (66.3)	495 (70.2)	1.00	ref				
Yes	85 (33.7)	210 (29.8)	0.95	0.85 - 1.07	0.412			
Smoking index								
Not currently	220 (87.3)	628 (89.1)	1.00	ref				
Currently	32 (12.7)	77 (10.9)	0.95	0.82 - 1.12	0.556			
Cough duration (days)[b]	30.0 (21.0 - 90.0)	30.0 (30.0 - 90.0)	1.00	1.00 - 1.00	0.658			
M. tuberculosis lineage								
Non-Beijing	207 (75.0)	540 (64.6)	1.00	ref				
Beijing	69 (25.0)	296 (35.4)	1.12	1.02 - 1.23	0.014	1.10	1.01 - 1.19	0.028
Index case highest smear grade								
Scanty1+	109 (39.5)	184 (22.0)	1.00	ref				
2+	77 (27.9)	236 (28.2)	1.20	1.04 - 1.38	0.011	1.10	0.96 - 1.26	0.151
3+	90 (32.6)	416 (49.8)	1.31	1.15 - 1.49	<0.05	1.23	1.09 - 1.39	0.001
Any drug resistance								
No	246 (89.1)	740 (88.5)	1.00	ref				
Yes	30 (10.9)	96 (11.5)	1.02	0.89 - 1.16	0.825			
Multidrug resistance								
No	266 (96.4)	816 (97.6)	1.00	ref				
Yes	10 (3.6)	20 (2.4)	0.88	0.70 - 1.11	0.296			
Cavities								
No cavity	138 (53.1)	334 (41.3)	1.00	ref				
Present	122 (46.9)	475 (58.7)	1.12	1.02 - 1.23	0.014	1.08	0.99 - 1.19	0.082

Extent x-ray disease[b]	40.0 (25.0 - 60.0)	45.0 (25.0 - 70.0)	1.00	1.00 - 1.00	0.075			
Age contact[b]	23.3 (12.5 - 41.4)	28.9 (15.1 - 45.6)	1.02	1.00 - 1.04	0.045	1.00	0.98 - 1.02	0.939
Contact sex								
Female	149 (54.0)	479 (57.3)	1.00	ref				
Male	127 (46.0)	357 (42.7)	0.97	0.90 - 1.04	0.35	0.97	0.88 - 1.05	0.433
Contact ethnicity								
Sunda	254 (92.0)	768 (91.9)	1.00	ref				
Other	22 (8.0)	68 (8.1)	1.00	0.98 - 1.02	0.933			
Sleep proximity								
Different room	217 (78.6)	586 (70.1)	1.00	ref				
Same room	59 (21.4)	250 (29.9)	1.11	1.03 - 1.19	<0.001	1.07	1.00 - 1.15	0.055
Hours with Case[b]	3.0 (1.0 - 7.0)	5.0 (2.0 - 10.0)	1.02	1.01 - 1.03	<0.001	1.02	1.01 - 1.03	<0.05
BCG vaccination								
No	35 (12.7)	197 (23.6)	1.00	ref				
Yes	241 (87.3)	639 (76.4)	0.86	0.80 - 0.92	<0.05	0.87	0.81 - 0.93	<0.05
Smoking								
Non Smoker	208 (75.4)	578 (69.2)	1.00	ref				
Current smoker	68 (24.6)	257 (30.8)	1.08	1.00 - 1.15	0.042	1.10	1.01 - 1.20	0.026
Diabetes[c]								
No diabetes	254 (92.0)	757 (90.5)	1.00	ref				
Pre-diabetes	13 (4.7)	53 (6.3)	1.07	0.94 - 1.22	0.282			
Diabetes	9 (3.3)	26 (3.1)	0.99	0.81 - 1.21	0.938			
BMI[b]	20.2 (17.0 - 24.3)	21.5 (17.8 - 25.2)	1.01	1.00 - 1.01	0.017	1.01	1.00 - 1.01	0.068

*Data are presented as N (%) unless indicated otherwise.
[a]Estimates obtained by multiple regression, adjusted for index case age, sex and smear grade, cavities and contact age, sex, sleep proximity to index case, hours spent with index case, smoking and body mass index.
[b]Median, interquartile range.
Abbreviations: PN: persistently negative; IGRA: interferon-gamma release assay; BCG: Bacillus Calmette–Guérin; IGRA: interferon-gamma release assay; APR: adjusted prevalence ratio calculated; 95% CI: 95% confidence interval.

Table S6.4. Association between BCG vaccination and IGRA at baseline and 14 weeks, according to index case lineage

Lineage	BCG vaccination	IGRA-negative N=275	IGRA-positive N=108	APR[a]	95% CI	P-value
Non-Beijing	No	22	131	1.00	ref	
	Yes	185	409	0.82	0.76 - 0.89	<0.001
Beijing	No	13	66	1.00	ref	
	Yes	56	230	0.96	0.86 - 1.06	0.4

[a]Estimates obtained by multiple regression, adjusted for index case age, sex and smear grade, cavities and contact age, sex, sleep proximity to index case, hours spent with index case, smoking and body mass index. Abbreviations: BCG: Bacillus Calmette–Guérin; IGRA: interferon-gamma release assay; APR: adjusted prevalence ratio; 95% CI: 95% confidence interval.

Table S6.5. Association between the index cases' *M. tuberculosis* lineage and IGRA conversion in household contacts who were IGRA negative at baseline (N=383)

	PN	Converter	RR	95% CI	P-value
Other isolates	206 (74.9)	69 (25.1)	1,00	ref	
Ancient Beijing	25 (56.8)	19 (43.2)	1,72	1,02 - 2,89	0,041
Modern Beijing	44 (68.8)	20 (31.3)	1,25	0,77 - 2,01	0,37

*Abbreviations: PN: persistently negative; RR: relative risk; CI: confidence interval.

References

1. Global Tuberculosis Report 2018. Geneva: World Health Organization.
2. Yates TA, Khan PY, Knight GM, Taylor JG, McHugh TD, Lipman M, et al. The transmission of *Mycobacterium tuberculosis* in high burden settings. The Lancet Infectious Diseases. 2016;16(2):227-38.
3. Middelkoop K, Mathema B, Myer L, Shashkina E, Whitelaw A, Kaplan G, et al. Transmission of Tuberculosis in a South African Community With a High Prevalence of HIV Infection. Journal of Infectious Diseases. 2015;211(1):53-61.
4. UNAIDS. Ending tuberculosis and AIDS: a joint response in the era of the Sustainable Development Goals – country submissions. Geneva, Switzerland; 2018 25 June 2018.
5. Cadena J, Castro-Pena NA, Javeri H, Hernandez B, Michalek J, Arzola AF, et al. Tuberculosis Patients Who Are A Potential Source for Unprotected Exposure in Health Care Systems: A Multicenter Case Control Study. Open Forum Infect Dis. 2017;4(4):ofx201.
6. Nebenzahl-Guimaraes H, van Laarhoven A, Farhat MR, Koeken VA, Mandemakers JJ, Zomer A, et al. Transmissible *Mycobacterium tuberculosis* Strains Share Genetic Markers and Immune Phenotypes. Am J Respir Crit Care Med. 2016.
7. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, et al. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. Nat Genet. 2013;45(10):1176-82.
8. Parwati I, van Crevel R, van Soolingen D. Possible underlying mechanisms for successful emergence of the *Mycobacterium tuberculosis* Beijing genotype strains. Lancet Infect Dis. 2010;10(2):103-11.
9. Mokrousov I, Narvskaya O, Otten T, Vyazovaya A, Limeschenko E, Steklova L, et al. Phylogenetic reconstruction within *Mycobacterium tuberculosis* Beijing genotype in northwestern Russia. Research in Microbiology. 2002;153(10):629-37.
10. Holt KE, McAdam P, Thai PVK, Thuong NTT, Ha DTM, Lan NN, et al. Frequent transmission of the *Mycobacterium tuberculosis* Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam. Nat Genet. 2018;50(6):849-56.
11. Merker M, Blin C, Mona S, Duforet-Frebourg N, Lecher S, Willery E, et al. Evolutionary history and global spread of the *Mycobacterium tuberculosis* Beijing lineage. Nat Genet. 2015;47(3):242-9.
12. Perez-Lago L, Campos-Herrero MI, Canas F, Copado R, Sante L, Pino B, et al. A *Mycobacterium tuberculosis* Beijing strain persists at high rates and extends its geographic boundaries 20 years after importation. Sci Rep. 2019;9(1):4687.
13. Guerra-Assuncao JA, Crampin AC, Houben RM, Mzembe T, Mallard K, Coll F, et al. Large-scale whole genome sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area. Elife. 2015;4.
14. Verrall AJ, G. Netea M, Alisjahbana B, Hill PC, van Crevel R. Early clearance of *Mycobacterium tuberculosis*: a new frontier in prevention. Immunology. 2014;141(4):506-13.
15. van Crevel R, Dockrell HM. TANDEM: understanding diabetes and tuberculosis. The Lancet Diabetes & Endocrinology. 2014;2(4):270-2.
16. Hill PC, Ota MO. Tuberculosis case-contact research in endemic tropical settings: design, conduct, and relevance to other infectious diseases. Lancet Infect Dis. 2010;10(10):723-32.
17. Lumb RVD, A.; Bastian, I.; Fitz-Gerald, M. Laboratory Diagnosis of Tuberculosis by Sputum Microscopy. Adelaide: South Australia Pathology; 2013.
18. Caviedes L, Lee TS, Gilman RH, Sheen P, Spellman E, Lee EH, et al. Rapid, efficient detection and drug susceptibility testing of *Mycobacterium tuberculosis* in sputum by microscopic observation of broth cultures. The Tuberculosis Working Group in Peru. J Clin Microbiol. 2000;38(3):1203-8.
19. Rathman G, Sillah J, Hill PC, Murray JF, Adegbola R, Corrah T, et al. Clinical and radiological presentation of 340 adults with smear-positive tuberculosis in The Gambia. Int J Tuberc Lung Dis. 2003;7(10):942-7.
20. Ralph AP, Ardian M, Wiguna A, Maguire GP, Becker NG, Drogumuller G, et al. A simple, valid, numerical score for grading chest x-ray severity in adult smear-positive pulmonary tuberculosis. Thorax. 2010;65(10):863-9.
21. QIAGEN. QuantiFERON-TB Gold ELISA Package Insert. 2017. Germantown, MD, USA. Available from: http://www.QuantiFERON.com.

22. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297-303.
23. Deatherage DE, Barrick JE. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. Methods Mol Biol. 2014;1151:165-88.
24. Coll F, McNerney R, Preston MD, Guerra-Assuncao JA, Warry A, Hill-Cawthorne G, et al. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Med. 2015;7(1):51.
25. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59(3):307-21.
26. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2019;35(3):526-8.
27. Coll F, Preston M, Guerra-Assuncao JA, Hill-Cawthorn G, Harris D, Perdigao J, et al. PolyTB: a genomic variation map for *Mycobacterium tuberculosis*. Tuberculosis (Edinb). 2014;94(3):346-54.
28. Shitikov E, Kolchenko S, Mokrousov I, Bespyatykh J, Ischenko D, Ilina E, et al. Evolutionary pathway analysis and unified classification of East Asian lineage of *Mycobacterium tuberculosis*. Sci Rep. 2017;7(1):9227.
29. Verrall AJ, Alisjahbana B, Apriani L, Novianty, Nurani A, van Laarhoven A, et al. Early clearance of *Mycobacterium tuberculosis*: the INFECT case contact cohort study in Indonesia. J Infect Dis. 2019; in press.
30. Walker TM, Ip CLC, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. The Lancet Infectious Diseases. 2013;13(2):137-46.
31. Behr MA, Verma S, Bhatt K, Lovey A, Ribeiro-Rodrigues R, Durbin J, et al. Transmission phenotype of *Mycobacterium tuberculosis* strains is mechanistically linked to induction of distinct pulmonary pathology. PLOS Pathogens. 2019;15(3):e1007613.
32. Abebe F, Bjune G. The emergence of Beijing family genotypes of *Mycobacterium tuberculosis* and low-level protection by bacille Calmette-Guerin (BCG) vaccines: is there a link? Clin Exp Immunol. 2006;145(3):389-97.
33. Kremer K, van-der-Werf MJ, Au BK, Anh DD, Kam KM, van-Doorn HR, et al. Vaccine-induced immunity circumvented by typical *Mycobacterium tuberculosis* Beijing strains. Emerg Infect Dis. 2009;15(2):335-9.
34. Anh DD, Borgdorff MW, Van LN, Lan NT, van Gorkom T, Kremer K, et al. *Mycobacterium tuberculosis* Beijing genotype emerging in Vietnam. Emerg Infect Dis. 2000;6(3):302-5.
35. de Jong BC, Hill PC, Aiken A, Awine T, Antonio M, Adetifa IM, et al. Progression to active tuberculosis, but not transmission, varies by *Mycobacterium tuberculosis* lineage in The Gambia. J Infect Dis. 2008;198(7):1037-43.
36. Marais BJ, Hesseling AC, Schaaf HS, Gie RP, van Helden PD, Warren RM. *Mycobacterium tuberculosis* Transmission Is Not Related to Household Genotype in a Setting of High Endemicity. Journal of Clinical Microbiology. 2009;47(5):1338-43.
37. Nebenzahl-Guimaraes H, Verhagen LM, Borgdorff MW, van Soolingen D. Transmission and Progression to Disease of *Mycobacterium tuberculosis* Phylogenetic Lineages in The Netherlands. J Clin Microbiol. 2015;53(10):3264-71.
38. Filliol I, Driscoll JR, van Soolingen D, Kreiswirth BN, Kremer K, Valetudie G, et al. Snapshot of moving and expanding clones of *Mycobacterium tuberculosis* and their global distribution assessed by spoligotyping in an international study. J Clin Microbiol. 2003;41(5):1963-70.
39. Cui Z, Lin D, Chongsuvivatwong V, Graviss EA, Chaiprasert A, Palittapongarnpim P, et al. Hot and Cold Spot Areas of Household Tuberculosis Transmission in Southern China: Effects of Socio-Economic Status and *Mycobacterium tuberculosis* Genotypes. Int J Environ Res Public Health. 2019;16(10).
40. Liu Y, Zhang X, Zhang Y, Sun Y, Yao C, Wang W, et al. Characterization of *Mycobacterium tuberculosis* strains in Beijing, China: drug susceptibility phenotypes and Beijing genotype family transmission. BMC Infect Dis. 2018;18(1):658.

41. Yang C, Shen X, Peng Y, Lan R, Zhao Y, Long B, et al. Transmission of *Mycobacterium tuberculosis* in China: a population-based molecular epidemiologic study. Clin Infect Dis. 2015;61(2):219-27.
42. Nebenzahl-Guimaraes H, Borgdorff MW, Murray MB, van Soolingen D. A novel approach - the propensity to propagate (PTP) method for controlling for host factors in studying the transmission of *Mycobacterium tuberculosis*. PLoS One. 2014;9(5):e97816.
43. Langlois-Klassen D, Senthilselvan A, Chui L, Kunimoto D, Saunders LD, Menzies D, et al. Transmission of *Mycobacterium tuberculosis* Beijing Strains, Alberta, Canada, 1991-2007. Emerg Infect Dis. 2013;19(5):701-11.
44. Coscolla M, Gagneux S. Consequences of genomic diversity in *Mycobacterium tuberculosis*. Semin Immunol. 2014;26(6):431-44.
45. Zhang Z, Lu J, Liu M, Wang Y, Qu G, Li H, et al. Genotyping and molecular characteristics of multidrug-resistant *Mycobacterium tuberculosis* isolates from China. J Infect. 2015;70(4):335-45.
46. Hanekom M, van der Spuy GD, Streicher E, Ndabambi SL, McEvoy CR, Kidd M, et al. A recently evolved sublineage of the *Mycobacterium tuberculosis* Beijing strain family is associated with an increased ability to spread and cause disease. J Clin Microbiol. 2007;45(5):1483-90.
47. Dixit A, Freschi L, Vargas R, Calderon R, Sacchettini J, Drobniewski F, et al. Whole genome sequencing identifies bacterial factors affecting transmission of multidrug-resistant tuberculosis in a high-prevalence setting. Sci Rep. 2019;9(1):5602.
48. Wang W, Hu Y, Mathema B, Jiang W, Kreiswirth B, Xu B. Recent transmission of W-Beijing family *Mycobacterium tuberculosis* in rural eastern China. Int J Tuberc Lung Dis. 2012;16(3):306-11.
49. Hanekom M, van der Spuy GD, Gey van Pittius NC, McEvoy CR, Ndabambi SL, Victor TC, et al. Evidence that the spread of *Mycobacterium tuberculosis* strains with the Beijing genotype is human population dependent. J Clin Microbiol. 2007;45(7):2263-6.
50. Hirsh AE, Tsolaki AG, DeRiemer K, Feldman MW, Small PM. Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. Proc Natl Acad Sci U S A. 2004;101(14): 4871-6.
51. Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, Narayanan S, et al. Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. Proc Natl Acad Sci U S A. 2006;103(8):2869-73.
52. van Laarhoven A, Mandemakers JJ, Kleinnijenhuis J, Enaimi M, Lachmandas E, Joosten LA, et al. Low induction of proinflammatory cytokines parallels evolutionary success of modern strains within the *Mycobacterium tuberculosis* Beijing genotype. Infect Immun. 2013;81(10):3750-6.
53. Krishnan N, Malaga W, Constant P, Caws M, Tran TH, Salmons J, et al. *Mycobacterium tuberculosis* lineage influences innate immune response and virulence and is associated with distinct cell envelope lipid profiles. PLoS One. 2011;6(9):e23870.
54. Tram TTB, Nhung HN, Vijay S, Hai HT, Thu DDA, Ha VTN, et al. Virulence of *Mycobacterium tuberculosis* Clinical Isolates Is Associated With Sputum Pre-treatment Bacterial Load, Lineage, Survival in Macrophages, and Cytokine Response. Front Cell Infect Microbiol. 2018;8:417.
55. Verrall AJ, Schneider M, Alisjahbana B, Apriani L, van Laarhoven A, Koeken V, et al. Early clearance of *Mycobacterium tuberculosis* is associated with increased innate immune responses. J Infect Dis. 2019; accepted for publication.
56. Floyd S, Ponnighaus JM, Bliss L, Warndorff DK, Kasunga A, Mogha P, et al. BCG scars in northern Malawi: sensitivity and repeatability of scar reading, and factors affecting scar size. Int J Tuberc Lung Dis. 2000;4(12):1133-42.

7

General discussion

General discussion

The two parts of this thesis are discussed separately, followed by an outlook on the present and future challenges and opportunities of whole genome sequencing of *M. tuberculosis*.

Part one: drug resistance

The World Health Organization recommends drug-susceptibility testing of *M. tuberculosis* complex (MTBC) isolates for all patients with tuberculosis to guide treatment decisions and improve outcomes. Still, it is estimated that only 29% of the 558,000 people (range 483,000 – 639,000) who developed multidrug/rifampicin-resistant tuberculosis in 2017 were detected and notified to national tuberculosis programmes[1]. These data come from 91 countries that have continuous surveillance systems, and 69 countries that rely on epidemiological surveys of bacterial isolates collected from representative samples of patients. In resource-limited settings where routine drug-susceptibility testing is not accessible to all tuberculosis patients owing to lack of laboratory capacity or resources, surveys conducted about every five years represent the most common approach to investigate the burden of drug resistance[1]. In the first part of my thesis I have shown that whole genome sequencing is not only a promising alternative to phenotypic drug susceptibility testing in the diagnosis of drug-resistant tuberculosis, but that it can also guide individualized clinical decision-making, and add to our understanding of the epidemiology of drug-resistant tuberculosis.

Performance of genotypic compared to phenotypic drug susceptibility testing

Whole genome sequencing is an ideal technique for drug-susceptibility prediction in *M. tuberculosis*, in which drug resistance is largely determined by chromosomal mutations, especially now the costs of sequencing have come down substantially and tools to facilitate the analysis of whole genome sequencing data are becoming increasingly available[2,3]. In **Chapter 2** of this thesis we showed that in a highly endemic setting in Indonesia the agreement of whole genome sequencing-based resistance prediction and phenotypic testing was high for isoniazid and rifampicin and lower for ethambutol and streptomycin. Recently, a large consortium showed, based on the analysis of more than 10,000 *M. tuberculosis* isolates collected from 16 countries across six continents and representing all major lineages, that whole genome sequencing can characterize susceptibility profiles to first-line antituberculosis drugs with high enough accuracy to be used in clinical practice[4]. The WHO targets of 90% sensitivity and 95% specificity for new molecular assays for *M. tuberculosis* were met

for all drugs, except the specificity for ethambutol. This study, like ours and most other studies, evaluated the performance of whole genome sequencing for drug susceptibility testing compared to phenotypic testing as the gold standard, but phenotyping is an imperfect standard[5,6], as we also showed in **Chapter 3**. Currently established procedures for *M. tuberculosis* phenotypic drug susceptibility testing classify clinical isolates as either drug-'resistant' or drug-'susceptible', on the basis of their ability to grow in the presence of a (mostly single) 'critical drug concentration'[6], defined as the lowest drug concentration that inhibits ≥95% of wild-type strains of bacilli that have not been exposed to the drug previously. This definition of drug susceptibility is based on laboratory tests, and patients with strains that are flagged resistant in this manner do not necessarily fail to respond to treatment[7]. In addition, breakpoint artefacts, where critical concentrations are set too high, have been reported as a major cause for systematic phenotypic drug susceptibility testing errors[5]. Our findings in **Chapter 3** that mutations associated with ethambutol and streptomycin resistance only moderately increased the minimum inhibitory concentration and that breakpoints were set too high, could explain the discrepant results between genotypic and phenotypic drug susceptibility testing for these drugs observed in **Chapter 2**. Therefore, discrepancies between genotypic and phenotypic drug susceptibility testing should be studied cautiously and not automatically be judged in favour of the phenotypic method.

Precision medicine

Individualization of therapy based on whole genome sequencing for the treatment of drug-resistant tuberculosis will require an extensive knowledgebase in which the effect of genomic mutations (and their combinations) on all relevant drugs is characterized. In addition, a clear understanding of what the molecular-level data are predicting is essential. Studies to date have tried to link mutations to minimum inhibitory concentrations[8,9], or to a probability of one of two outcomes - phenotypic resistance or susceptibility[10] for certain drugs. In **Chapter 3** we linked whole genome sequence-determined mutations to minimum inhibitory concentrations for all first- and most of the second-line drugs and found that different mutations lead to different levels of resistance; knowing the underlying mutations can guide clinical decision-making and facilitate therapeutic drug monitoring. The ultimate aim would be for molecular data to predict treatment outcome directly. The clinical implications of some mutations known to lead to high-level isoniazid (*katG* S315T), rifampicin (*rpoB* S450L) and fluoroquinolone (*gyrA* 94) resistance and treatment failure have been outlined in a consensus statement[11], and are in line with our findings. The data we presented in **Chapter 3** furthermore indicated that particular mutations in *rpoB* only cause a slight increase in the minimum inhibitory concentration for rifampicin and might justify the prescription of high-dose rifampicin in patients with these mutations. Similarly, a

recent paper by Farhat *et al.* discovered that promotor mutations have smaller average effects on resistance than gene body mutations[9] and may be overcome by increased dosing. This way, whole genome sequencing can ultimately be used to personalize tuberculosis treatment, based on the detailed characterisation of the resistance profile of the bacterium[12] (**Figure 7.1**). HIV has already set an example, as nowadays genotypic resistance testing in developed countries is performed routinely as companion diagnostics in HIV therapy[13].

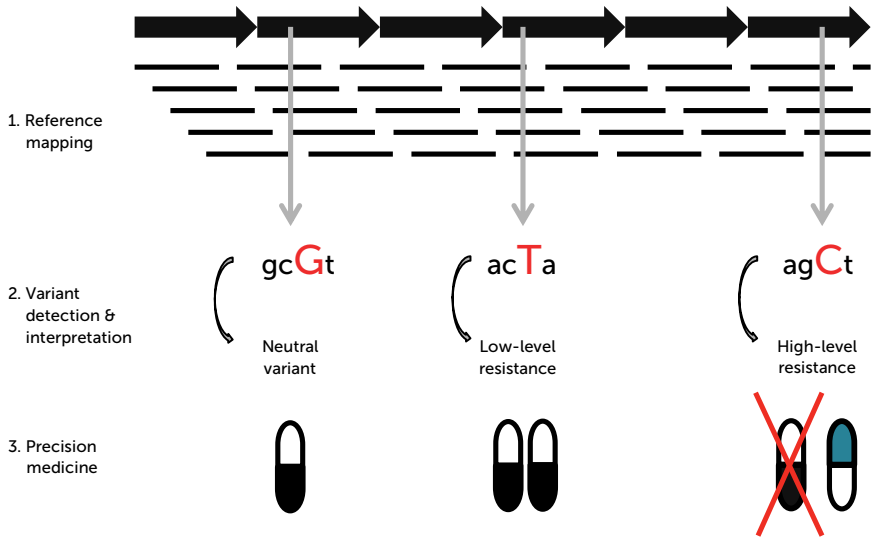


Figure 7.1. Drug resistance mutations can guide clinical decision-making. "Neutral variant" indicates that the mutation has no effect on the minimum inhibitory concentration for a specific drug.

Implementation in routine clinical diagnostics

Sequencing-based drug susceptibility testing offers other benefits for diagnosis and therapy besides the prospect of personalizing tuberculosis treatment. First, if all resistance-conferring mutations were known, this would facilitate individualized treatment by determining which drugs to give, not only which ones to avoid. Second, as drug resistance can be predicted from the genome sequence before phenotypic results become available, treatment with an appropriate regimen can be initiated earlier, reducing the risk of treatment failure, resistance amplification and transmitting drug-resistant tuberculosis. The potential of whole genome sequencing being performed directly on clinical samples would reduce the time to appropriate treatment initiation even further. Walker and colleagues[14] have shown that data from whole

genome sequencing could be used clinically, to predict first-line drug resistance, drug susceptibility or to identify drug phenotypes that cannot yet be genetically predicted, and that this approach could be integrated into routine diagnostic workflows. Data on the routine use of genetic sequencing for antituberculosis drug resistance are scarce and only from industrialised countries. However, whole genome sequencing should be implemented particularly in low-income, high-burden settings where many still rely on empirical treatment regimens and the need is greatest, to actually revolutionize the diagnosis of drug-resistant tuberculosis. The studies that are part of this thesis are all performed in highly endemic, limited-resource countries, but whole genome sequencing was applied retrospectively, for research purposes. Research on the implementation of genome sequencing for routine drug-susceptibility testing in resource-limited settings has not been performed and is urgently needed.

Further study on possible loss-of-fitness-compensating mutations

Whole genome sequencing has facilitated significant progress in understanding the complex biology and epidemiology of drug-resistant tuberculosis, next to the clinical benefits of *M. tuberculosis* sequencing-based drug susceptibility testing. Mutations that lead to resistance can be deleterious or produce a fitness cost in the absence of the drug. One could therefore hypothesise that resistant strains are less virulent or less easily transmitted than drug-sensitive strains[15-17]. However, recent work has shown that additional mutations often follow or coincide with drug resistance mutations, and that these mutations can compensate for deleterious effects, restoring their initial growth capacity[18-20]. Based on the data In **Chapter 4** one may hypothesize that less fit drug-resistant *M. tuberculosis* strains are more likely to cause active disease among people living with diabetes, because of their lower host defence[21], and that these strains would have less mutations compensating for the resistance-associated loss of fitness. Our data did not support this hypothesis: we found no differences in the presence of potential rifampicin resistance-compensating mutations between strains infecting diabetes and non-diabetes patients.

A large number of putative compensatory mutations have been reported in the *rpoA*, *rpoB* and *rpoC* genes encoding the α-, β- and β'-subunits of the RNA polymerase[22,23]. However, except for some notable exceptions[24], the evidence that they compensate for a loss of fitness caused by the primary resistance mutation has mainly been statistical[23,25,26]. Therefore, in addition to the results presented in **Chapter 4**, where we focused on all non-synonymous mutations in *rpoA, rpoB* and *rpoC*, we sought to identify which of these mutations co-localized onto the 3D structure of the *Mycobacterium smegmatis* RNA polymerase. To this end, we counted the number of *rpoA/B/C* mutations in *M. tuberculosis* isolates with a known rifampicin-resistance mutation that converged in the 3D structure of the molecule as evidence for their role

in compensation for the loss of fitness caused by the resistance mutation. Conceptually this takes the approach of counting the number of "convergent" mutations that occur a the same location in the sequence a step further by examining it in the 3D structure and by putting it on a formal statistical basis (schematically represented in **Figure 7.2**). We projected the mutations that co-occur with the various rifampicin resistance mutations (separately for *rpoB* S450L, V170F and for *rpoB* mutations at position 435 and 445) onto the structure of the *M. smegmatis* RNA polymerase, which is 91% identical at the sequence level with *M. tuberculosis*[27]. We used the tool mutation3D[28] that uses randomizations of the locations of mutations in the 3D structure to analyse whether certain regions are overrepresented, and explicitly includes convergent mutations in its analysis. For *rpoB* S450L we found one, highly significant (E < 1E-5) cluster that contains one mutation in *rpoA* (V183A) and five mutations in *rpoC* (V483G, V483A, W484S, I491V and V517L), distributed over two regions in the sequence (**Figure 7.3**). It should be noted that mutations in this region have been associated with resistance mutations before[22,29] although not with all specific mutations that we found. For the other rifampicin resistance mutations we observed no significant clusters of compensating mutations in the 3D structure. The co-localizing mutations in the significant cluster occurred in rifampicin-resistant isolates from three tuberculosis patients with, and six without diabetes, and there was no difference in the presence of these mutations between strains infecting diabetes and non-diabetes patients.

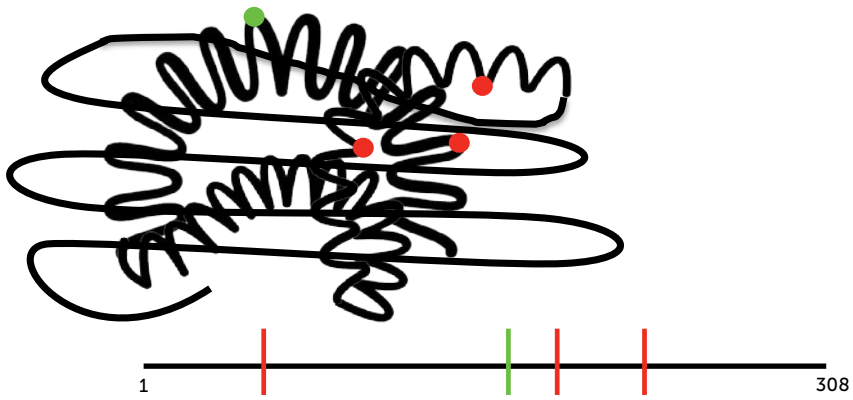


Figure 7.2. Missense mutations in a hypothetical protein. The linear protein model is given below the 3D structure to illustrate the importance of studying co-localisation of mutation in 3D structures. The mutations in red co-localize in the 3D model, whereas the mutation in green seems to cluster with the two red mutations in the linear protein model but falls outside the cluster in the 3D structure.

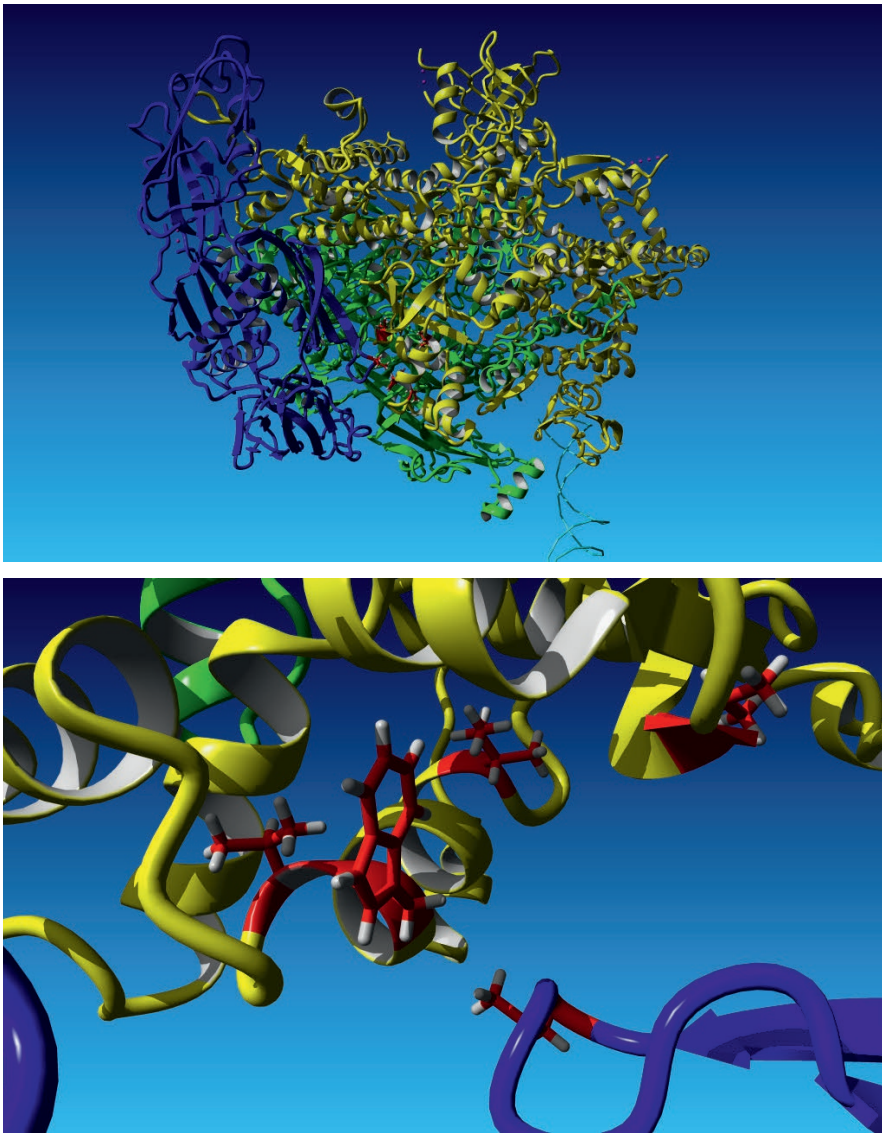


Figure 7.3. Potential rifampicin resistance-compensating mutations in *rpoA* and *rpoC* that co-localize on the 3D structure of *M. smegmatis* RNA polymerase. Mutations identified in *M. tuberculosis* isolates with rifampicin resistance mutations were mapped onto the RNA polymerase structure of *M. smegmatis*. The co-localised mutations (residues coloured in red) fall at the interface of *rpoA* (α- subunit, dark blue) and *rpoC* (β'-subunit, yellow) of the *M. smegmatis* RNA polymerase. *RpoB* (β-subunit) is coloured in green.

Although the clustered mutations were not associated with the diabetes phenotype, our findings add evidence to the possible compensatory role of these mutations. Their interaction with the strain genetic background and the fitness cost of the resistance mutations may together influence the spread and maintenance of resistant variants in the population, a process called epistasis[30] (**Figure 7.4**).

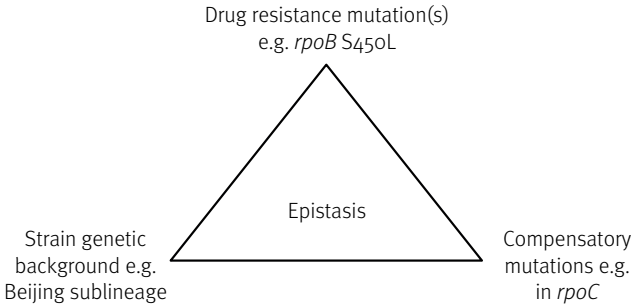


Figure 7.4. Epistatic interactions: the effect of drug resistance mutations alone or in combination as well as the presence of compensatory mutations affects the fitness of drug-resistant *M. tuberculosis* strains as a function of the strain genetic background.

Based on the findings in this thesis, and those of others, I think that whole genome sequencing-based drug susceptibility testing of *M. tuberculosis* can drastically improve the diagnosis of drug-resistant tuberculosis, but that there are important challenges to be addressed (**Box 7.1**).

Box 7.1. Conclusions and challenges of using whole genome sequencing for drug susceptibility testing

Conclusions

- Whole genome sequencing-based drug susceptibility testing of *M. tuberculosis* has the potential to drastically improve the diagnosis of drug-resistant tuberculosis:
  - Faster and more accurate diagnosis
  - Possibility to predict drug susceptibility in addition to resistance
- By generating whole genome sequencing data and linking these to relevant clinical parameters using state-of-the-art machine learning techniques, we can:
  - Increase our understanding of the molecular basis of antibiotic resistance
  - Reach better performance for several drugs while significantly reducing the number of candidate mutations to consider[31,32]
  - Design PCR assays that target a small but optimal number of markers

Challenges

- Implementing whole genome sequencing in high-burden countries:
  - Need for expensive equipment and highly trained personnel
  - Requires complex bioinformatic procedures
- Lack of understanding of the genetic basis of antibiotic resistance complicates the interpretation of whole genome sequencing data

Part two: tuberculosis disease phenotype and transmission

The second part of this thesis focused on the link between *M. tuberculosis* genotype and tuberculosis disease phenotypes.

Exposure to *M. tuberculosis* can be followed by early clearance through innate immunity, direct development of active disease, or latent infection that may or may not reactivate up to several decades after initial exposure. Traditionally, these different phenotypes have been attributed to host and environmental variables[33]. Because of the limited genetic diversity within MTBC (99.9% nucleotide identity[34]) compared to other bacteria, the view has been that this "negligible" strain-to-strain genetic variation had no phenotypic relevance, i.e. lineage- or strain-specific differences in disease

manifestations. However, recent advances in whole genome sequencing of large MTBC clinical strain collections from global sources have revealed more genomic diversity than previously anticipated. In addition to the genomic diversity across MTBC clinical strains, findings from many experimental studies have led to a change in paradigm by demonstrating the phenotypic impact of this genetic diversity. For example, studies have reported differences between clinical strains with respect to their transcriptomic profiles[35,36], protein and metabolite levels[36,37], methylation profiles[37,38], drug susceptibility[39] and cell wall structure[40-42]. In addition, MTBC genetic diversity has also been shown to influence disease severity and human to human transmission, with "modern" lineages showing a faster progression to disease and shorter latency periods compared to strains from the "ancestral" clades[43-46]. At the lineage level, the East-Asian lineage (Lineage 2) has been linked to increased transmission and extrapulmonary disease[47,48] compared to the other lineages.

Associations beyond *M. tuberculosis* lineage level

However, associations reported between lineage and disease phenotype are not consistent[46] and considerable intra-lineage diversity exists. For example, Lineage 2 strains exhibit a range of inflammatory and virulence phenotypes[49,50]. This could be due to the fact that Lineage 2 consists of a variety of sub-lineages, including various variants of "Beijing" strains, which might differ phenotypically[42]. Beijing strains have been proposed to possess selective advantages comprising an increased capacity to acquire drug resistance, increased transmissibility, hypervirulence and more rapid progression to disease after infection[51]. Still, also the association of Beijing strain infection with disease manifestations is not consistent[52]. This heterogeneity, together with evidence for higher virulence of modern Beijing strains in comparison with ancestral sublineage strains[53,54], suggests the existence of inter-strain variation within sublineages affecting the disease phenotype, or an important role for the interaction between the host and pathogen genotypes (co-evolution)[55].

In general, the genetic factors that are used to classify *M. tuberculosis* in phylogenetic groups, whether at supra-lineage, lineage, or sub-lineage level, do not necessarily provide clear genetic determinants of tuberculosis disease phenotype. For example, the genes that have been found to be associated with the invasion or survival of *M. tuberculosis* in the central nervous system[56], important for the development of tuberculous meningitis, can be found in all lineages of *M. tuberculosis*. Detecting the *M. tuberculosis* genetic determinants of phenotypic differences requires analytical methods that index genomic diversity in a more comprehensive manner. Many genotyping schemes have been developed for *M. tuberculosis* in the past[57], but only whole genome sequencing provides information about the whole genome, and can identify virtually all varieties of markers detected by other genotyping methods. It is

therefore highly accurate and precise in detecting variability among strains and provides a wealth of information at every level possible. In **Chapter 5**, we did not find an association between *M. tuberculosis* lineage and the tuberculous meningitis phenotype, but using whole genome sequencing we did find evidence for homoplastic genetic variation in three genes that were associated with the disease phenotype. The one study published on this topic after our study used a similar approach, and also identified variants – though different from the ones we found - associated with tuberculous meningitis in a lineage-independent fashion[58].

Challenges associated with microbial genome-wide association studies

The high resolution of whole genome sequencing also comes with analytical challenges in the quest to link genotype to phenotype. Instead of two categories, when linking ancient vs. modern lineages to differences in disease phenotype, the number of possible variations increases tremendously when examining all possible SNPs, insertions, and deletions across the entire genome, even in a clonal organism as *M. tuberculosis*. Whole genome sequencing generates thousands of potentially relevant mutations, the large majority of which will be noise that is not related to the phenotype of interest. In human genetics, genome-wide association studies (GWAS) have been developed and methods advanced significantly in the past fifteen years. The purpose of GWAS is to identify statistically significant associations that may indicate the presence of a causal relationship between genotype and phenotype while rejecting spurious associations arising from confounding factors. Compared to human GWAS, smaller genome sizes and the ability to manipulate the genomes in the laboratory may improve the power and computational ease of bacterial GWAS and facilitate the confirmation of candidate loci[59]. Bacterial association mapping nonetheless remains a technical challenge in search of an optimal methodological approach. Microbial GWAS must overcome various confounding factors, such as the stronger population stratification[60], and the selection force on bacteria-related phenotypes. The problem of population stratification is especially critical in highly clonal (rarely recombining) bacteria, and in those with separate geographic- or host-associated subpopulations, such as *M. tuberculosis*[59].

Phylogenetic approach

Fortunately, population stratification and selection enable the adoption of a phylogenetic solution in microbial GWAS. Phylogenetic trees allow for the detailed identification of genetic relationships, not only at the level of population clusters, but also at the resolution of subpopulations and individual relationships. By combining a signal of selection with phenotypic associations, the phylogeny helps to decipher if an association is due to genetic relatedness rather than causality for the phenotype of interest (e.g. convergent mutations that occur only in cases and not in controls are

more likely to be causally related to the phenotype of interest). In **Chapter 5**, we used this 'homoplasy counting' method to find *M. tuberculosis* genetic variations associated with the tuberculous meningitis phenotype by focusing on repeated and independently emerging mutations occurring more often on branches of isolates from tuberculous meningitis patients compared to those from pulmonary tuberculosis patients (**Figure 7.5**).

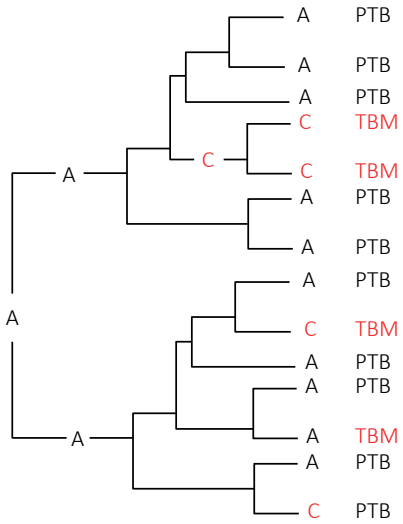


Figure 7.5. The homoplastic A→C mutation occurs independently in the tree and is associated with the tuberculous meningitis (TBM) phenotype.

The pitfall in using the phylogeny to adjust for associations due to population stratification ('phylogenetic noise') is the risk of masking causal variants, because differences between lineages account for large proportions of both phenotypic and genotypic variability. Correcting for lineage-effects will lead to overcorrection when these effects are biologically meaningful. In **Chapter 6**, we discovered that *M. tuberculosis* isolates belonging to the Beijing sublineage were associated with increased transmission. Additional analyses, including host-related variables, showed that this increased transmissibility was related to evasion of BCG-protection by these "Beijing" isolates. In this case, the observed lineage effect could, in a next step, guide the identification of loci that contributed to the significant lineage-level association. Disentangling the effects of a single variant from those related to lineage is potentially challenging, but has been shown to increase the power of microbial GWAS[61]. This strategy provides an alternative to prioritizing variants based solely on locus-specific significance, but it carries risks, because lineage-associated effects are more

susceptible to confounding with population-stratified differences in environment or sampling. Hence, dealing with population stratification represents a trade-off between the power to detect genuine associations of population-stratified variants and robustness to unmeasured, population stratified confounders.

Correction for multiple testing

Apart from the mentioned confounding factors, the other major source of false-positive associations is the multiple testing-problem that is intrinsic to GWAS. Various dimension-reduction approaches can be used to counter this problem. First, well-known methods to adjust for the large number of tests are the Bonferroni correction or the Benjamini-Hochberg procedure. Second, one can use independent datasets and test the most significant associations from the one set in the other. In **Chapter 5**, we used this approach and identified a list of top variants that were most likely associated with the tuberculous meningitis phenotype, and tested only these in a validation set of independent isolates, reducing the number of tests to correct for. A third method is the burden test, to aggregate information across several variant sites within a gene to enrich association signals and to reduce the penalty of multiple testing, which we also applied in **Chapter 5**. We also tested variants at pathway-level, an even higher level of aggregation. Fourth, a polygenic association test can be employed when one assumes that the trait in question is affected by a large number of variants on different genes. Fifth, one may choose to filter rare variants occurring in less than 1% of the isolates. Due to the highly clonal population structure of *M. tuberculosis*, many variants are rare and only occur in a very small fraction of the isolates (<0.01), increasing the number of variants to test for. Recent studies have found that 75% of the variants were found at a frequency of less than 0.0028 and 0.00054 in lineages 2 and 4, respectively in 4,408 sequenced genomes[62], and that 83% of the variants had a frequency of ≤0.01 in 1,452 genomes[9]. Depending on the underlying assumption whether the cumulative effect of these rare variants together could explain a large amount of the variance in risk, they can be included in a burden test, a polygenic association test, or filtered-out to reduce the multiple testing issues.

Validation

We know from the experience of human GWAS that some loci found to be associated with a trait (after correction for multiple testing), can turn out to have little or no functional significance[63]. Therefore, unless the associated locus has been previously shown to affect the phenotype, functional validation is desirable. The strongest evidence for the inference of functional associations remains experimental validation in the laboratory. The genetic variants we identified in **Chapter 5 and 6**, associated with tuberculous meningitis and transmissibility, respectively, could be validated experimentally. For example in a stimulation experiment where human macrophages

from tuberculosis patients are infected with *M. tuberculosis* strains with and without the identified mutations, a method that has been applied successfully in our laboratory before[64]. Even stronger proof for a functional role would come from validation by genetic disruption and reconstruction of *M. tuberculosis* strains, coupled with loss and regain of the phenotype, using transposon insertion mutagenesis[65]. With this 'top down' approach, findings from hypothesis-free GWAS can be used to generate hypotheses to be tested either in hypothesis-driven genetic association studies or experimentally in the laboratory.

Whole genome sequencing of *M. tuberculosis* has already, and will continue to improve our understanding of the evolutionary and molecular processes involved. Still, truly understanding the biological mechanisms underlying particular tuberculosis disease phenotypes studied in this thesis and by others remains extremely difficult. I believe that one of the major challenges is to define the phenotypic impact of specific *M. tuberculosis* genetic variants in the context of other bacterial determinants, the *M. tuberculosis* strain's genetic background and other host and environmental factors also impacting the phenotype. This requires carefully considered study designs with an analytical approach integrating all these factors.

Future perspectives

Understanding the *M. tuberculosis* genome biology is a necessary step towards the control of tuberculosis. The *M. tuberculosis* genome was first sequenced in 1998 and thanks to progress in high-throughput sequencing and decreasing raw sequencing costs many more *M. tuberculosis* genomes have since been sequenced[66-68], including the more than 1,000 in this thesis. The community consensus is that *M. tuberculosis* whole genome sequencing is now advanced enough to inform clinical decisions and public health policy[69]. Several challenges associated with the standardization, validation and implementation of *M. tuberculosis* whole genome sequencing remain to be addressed (**Table 7.1**) that will require political commitment and the participation of supranational laboratories and regulatory authorities. In addition, the research community as a whole should continue to improve the technical and analytical aspects of whole genome sequencing. I would encourage health policy makers to commit resources to ensure access to standardized and validated whole genome sequencing methodologies, especially in high-burden countries, where they will have the largest impact on tuberculosis control.

Table 7.1. Overview of challenges and opportunities associated with applications of *M. tuberculosis* whole genome sequencing discussed in this thesis

Challenges	Opportunities
General	
! High costs	+ Information on the entire genome
! Data storage capacity	+ High resolution
! Bioinformatics support & IT infrastructure	+ PacBio sequencing: repetitive regions and methylation sites[71]
! Difficult to sequence regions with repetitive elements	+ Can generate hypotheses to be tested experimentally
! Maintenance of software tools[70]	
Drug susceptibility testing	
! Implementation in low-resource settings	+ Faster than phenotypic DST
! Relies on quality of mutation database	+ Personalized medicine
! Resistance mutation catalogues rely mostly on data from lineages 2 & 4[72,73]	+ Pipeline adaptation for NTM
	+ Resistance to many drugs at no extra costs
! Mixed infections or heteroresistance[74]	+ Detection of novel resistance mutations[9]
! Incomplete understanding of the genetic basis of *M. tuberculosis* drug resistance	+ Sequencing directly on patient samples will provide results faster[76]
! Standardisation of methodology and genotype nomenclature[75]	+ Mapping transmission routes of drug-resistant tuberculosis
! Implementation in routine clinical diagnostics	+ Input for designing PCR-based assays
	+ Easy-to-use tools for raw sequencing files
Linking genotype to phenotype	
! Confounding by host-related factors	+ Phylogeny-based methods to correct for population stratification and selection[78]
! Strong population stratification[77]	
! Multiple testing problem	+ Possibilities for dimension reduction (e.g. polygenic and burden tests)
! Many rare variants	
! Validation	+ Genome-to-genome approach[79]

Abbreviations: DST: drug susceptibility testing; NTM: nontuberculous mycobacteria; PCR: polymerase chain reaction.

Implementation in routine clinical diagnostics

Whole genome sequencing has huge potential for applications in diagnostic microbiology including isolate characterization, drug susceptibility testing and establishing the source of recurrent infections and between-person transmissions. All of these have obvious clinical relevance and could provide additional information or even replace the knowledge obtained through standard clinical microbiology techniques. However, to date, *M. tuberculosis* whole genome sequencing has been mostly used for research purposes. One major limitation to rapidly obtaining useful information in a clinical setting is that analysis pipelines for microbial genomics have generally been

developed for fundamental research or public health epidemiology[80]. This usually means that the pipeline consists of a very thorough and sophisticated workflow with a large number of options and parameters. While this is desirable from a researcher's perspective, it is clearly problematic for real-time analysis in a clinical setting. The user requires in-depth knowledge about the purpose each tool serves, the relative strengths and weaknesses of each approach, and a functional understanding of the important parameters. Furthermore, knowledge of the Linux operating system and command line is often essential for most of the advanced bioinformatic analyses; something clinical microbiologists are rarely trained for. Hence, there is a need for simpler software tools, free of charge, able to analyse data directly from the initial raw sequencing files, providing the basic functionalities all in one. Fortunately, online software tools are becoming increasingly available to support the clinical microbiologist in the analysis of *M. tuberculosis* whole genome sequencing data[2,81] (**Table 7.2**).

Table 7.2. 'Easy-to-use' non-commercial genome analysis tools using raw sequencing data as input

	Year of introduction	Web-based	Batch mode	Variants reported	Number of drugs tested
CASTB[82]	2015	Yes	No	None	6
KvarQ[83]	2014	No	Yes	Only resistance mutations	9
Mykrobe Predictor TB[84]	2015	No	In command line version only	Only resistance mutations	12
PhyResSE[85]	2015	Yes	Yes	All mutations	12
TBProfiler[72]	2015	Yes	In command line version only	All mutations in candidate genes	12
TGS-TB[86]	2015	Yes	Yes	All core genome mutations	8

The implementation of whole genome sequencing in routine clinical diagnostics is particularly promising for drug susceptibility testing. The prediction of drug susceptibility based on drug resistance mutations has been shown reliable for first-line drugs[4,87], especially in settings with low drug resistance rates, and Public Health England has already decided to stop phenotypic drug susceptibility testing of isolates that are predicted to be genotypically susceptible to all first-line drugs. Similar decisions have been made in the Netherlands (planned for the end of 2019) and New York[4]. Ideally, genetic data should be linked to minimum inhibitory concentrations and, preferably, to patient outcomes. The aim would be to have a comprehensive,

standardized and curated resource for clinically relevant genetic mutations and associated metadata (geographic location, minimum inhibitory concentration, response to treatment)[88], similar to the in 2007 established Stanford HIV Drug Resistance Database[89]. Whole genome sequencing on a large number of strains collected worldwide, coupled with these metadata can provide the appropriate statistical power to identify the subset of mutations predictive of treatment failure to any given drug. Furthermore, linking this with therapeutic drug monitoring results would enable clinicians to determine if treatment fails because of an increased minimum inhibitory concentration or because of a low blood concentration.

Understandably, efforts to develop whole genome sequencing-based microbial diagnostics have focussed on high-resource settings. Studies on the impact and cost effectiveness of routine whole genome sequencing in high burden settings are needed to determine the feasibility of whole genome sequencing in this setting. Currently, most hospitals in developing countries do not even benefit from a clinical microbiology laboratory. This presents an obvious challenge, but at the same time an outstanding opportunity for low- to middle-income countries to get up to speed with whole genome sequence-based developments in real-time clinical diagnostics, rather than adopting classical microbiological phenotypic drug susceptibility testing that may eventually be phased out in high-income countries. The installation of a molecular laboratory based around a benchtop sequencer might be an ideal investment, as it is neither far more expensive nor more complex than setting up a standard clinical microbiology laboratory. The introduction of easy-to-use software tools that predict drug resistance profiles based on raw sequence input files has already brought the implementation of sequence-based drug susceptibility testing in these settings a step closer. A possible alternative or intermediate solution to bridge the gap in drug susceptibility testing between high- and low-income countries is by translating the drug resistance markers discovered by whole genome sequencing studies into the development of new, improved molecular assays that test drug resistance based on these markers and are more affordable and easier to use and implement.

A thus far relatively unexplored area is the application of whole genome sequencing in routine clinical diagnostics of nontuberculous mycobacteria (NTM), several species of which are increasingly recognized as important opportunistic pathogens of humans[90]. Despite the availability of whole genome sequencing technologies, limited effort has been put into the genetic characterization of NTM species. This year, we have started a project in the medical microbiology department to sequence all clinical NTM isolates and develop an automated bioinformatics pipeline for species identification and genotypic drug susceptibility testing. Both steps are crucial to guide treatment[91]; using whole genome sequencing can significantly speed up this process, leading to

more rapid diagnosis and streamlined treatment. Doing this in parallel with minimum inhibitory concentration determination will also enable the identification of novel resistance markers.

Linking genotype to phenotype

Future genetic association studies linking *M. tuberculosis* genotype to phenotype should include the human genotype as a variable of interest, because the global phylogeographic structure of *M. tuberculosis* and the association between *M. tuberculosis* strains and their human host populations provides solid evidence that these are closely related[92-94]. The interaction of the *M. tuberculosis* and human genotype might provide insight in the way the different *M. tuberculosis* lineages have adapted to specific human populations[55]. Associations between particular human SNPs and specific *M. tuberculosis* genotypes[48,95] support the concept of 'co-evolution' of *M. tuberculosis* and the human host. Such 'genome-to-genome' analyses have a great potential to uncover relevant interactions between the pathogen and host genome and their joint effects on phenotypes[79]. For example, *M. tuberculosis* genetic variation *X* could have a different effect in a patient with genotype *Y*, than in a patient with genotype *Z*. The success of this approach has recently been shown in pneumococcal meningitis[96]. We would like to use this method in follow-up studies, given that we have the availability of DNA samples for many of the patients who were included in the studies in this thesis.

The genes we found to be possibly related to the meningeal disease manifestation, and the phylogenetic association with a higher transmission rate through 'BCG-escape', need to be validated in follow-up studies. A genetic association does not prove causality, even though we adjusted for host-related factors and confounding by *M. tuberculosis* population stratification. To confirm that the found associations are functionally/biologically relevant, the next step would be to test the associations in another patient cohort, in a hypothesis-driven genetic association study, or experimentally in the laboratory. The advantage of functional validation in the laboratory is that conditions can be controlled in a way that other factors affecting the phenotype of interest are removed, kept constant, or varied on purpose. For instance to validate the genetic variants found to be associated with the tuberculous meningitis phenotype, macrophage infection models could be used to measure gene expression *in vitro* by infecting healthy macrophages with various *M. tuberculosis* strains with and without the identified mutations. In addition, macrophages from pulmonary tuberculosis and tuberculous meningitis patients could be infected with standard *M. tuberculosis* strains, or pulmonary tuberculosis and tuberculous meningitis patient macrophages could be infected with their 'own' *M. tuberculosis* strains to help disentangle the effect of *M. tuberculosis* or host genotype, or their interaction.

Functional genetics

Genetics does not "say it all". In the end, the work described in this thesis can be used as a basis to work towards 'functional genetics' research, linking the *M. tuberculosis* genome to *M. tuberculosis* transcription, transcriptional regulation (e.g. epigenetics), lipidomics, proteomics, and/or metabolomics, in appropriate models (e.g. macrophages or blood or cerebrospinal fluid of healthy individuals or tuberculosis patients). Integrating multiple types of these quantitative molecular measurements creates a more holistic understanding of the functional relevance/consequences of *M. tuberculosis* genetic variation. It could also shed light on the biological role of *M. tuberculosis* genes, as the functions of 27% of the encoded proteins in the annotated version of the H37Rv strain have yet to be determined[97] while the functions of the remaining proteins have been based mostly on sequence comparison[98].

M. tuberculosis whole genome sequencing can be applied to the entire spectrum from molecule to man to population and will help to find answers to fundamental research questions, which can be translated in improved diagnostics and patient care, to ultimately be adopted for tuberculosis surveillance and control. I hope the findings in this thesis will help to develop better tools and strategies to control a disabling disease, for which the burden largely lies in resource-poor countries.

References

1. Global Tuberculosis Report 2018. Geneva: World Health Organization.
2. Faksri K, Tan JH, Chaiprasert A, Teo YY, Ong RT. Bioinformatics tools and databases for whole genome sequence analysis of *Mycobacterium tuberculosis*. Infect Genet Evol. 2016;45:359-68.
3. Ngo TM, Teo YY. Genomic prediction of tuberculosis drug-resistance: benchmarking existing databases and prediction algorithms. BMC Bioinformatics. 2019;20(1):68.
4. Consortium CR, the GP, Allix-Beguec C, Arandjelovic I, Bi L, Beckert P, et al. Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing. N Engl J Med. 2018;379(15):1403-15.
5. Heyckendorf J, Andres S, Koser CU, Olaru ID, Schon T, Sturegard E, et al. What Is Resistance? Impact of Phenotypic versus Molecular Drug Resistance Testing on Therapy for Multi- and Extensively Drug-Resistant Tuberculosis. Antimicrob Agents Chemother. 2018;62(2).
6. Bottger EC. The ins and outs of *Mycobacterium tuberculosis* drug susceptibility testing. Clin Microbiol Infect. 2011;17(8):1128-34.
7. Canetti G, Froman S, Grosset J, Hauduroy P, Langerova M, Mahler HT, et al. Mycobacteria: Laboratory Methods for Testing Drug Sensitivity and Resistance. Bull World Health Organ. 1963;29:565-78.
8. Jamieson FB, Guthrie JL, Neemuchwala A, Lastovetska O, Melano RG, Mehaffy C. Profiling of *rpoB* mutations and MICs for rifampin and rifabutin in *Mycobacterium tuberculosis*. J Clin Microbiol. 2014;52(6):2157-62.
9. Farhat MR, Freschi L, Calderon R, Ioerger T, Snyder M, Meehan CJ, et al. GWAS for quantitative resistance phenotypes in *Mycobacterium tuberculosis* reveals resistance genes and regulatory regions. Nat Commun. 2019;10(1):2128.
10. Miotto P, Tessema B, Tagliani E, Chindelevitch L, Starks AM, Emerson C, et al. A standardised method for interpreting the association between mutations and phenotypic drug resistance in *Mycobacterium tuberculosis*. Eur Respir J. 2017;50(6).
11. Dominguez J, Boettger EC, Cirillo D, Cobelens F, Eisenach KD, Gagneux S, et al. Clinical implications of molecular drug resistance testing for *Mycobacterium tuberculosis*: a TBNET/RESIST-TB consensus statement. Int J Tuberc Lung Dis. 2016;20(1):24-42.
12. Cox H, Hughes J, Black J, Nicol MP. Precision medicine for drug-resistant tuberculosis in high-burden countries: is individualised treatment desirable and feasible? The Lancet Infectious Diseases. 2018; 18(9):e282-e7.
13. Lengauer T, Pfeifer N, Kaiser R. Personalized HIV therapy to control drug resistance. Drug Discov Today Technol. 2014;11:57-64.
14. Walker TM, Kohl TA, Omar SV, Hedge J, Del Ojo Elias C, Bradley P, et al. Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study. The Lancet Infectious Diseases. 2015;15(10):1193-202.
15. Sander P, Springer B, Prammananan T, Sturmfels A, Kappler M, Pletschette M, et al. Fitness cost of chromosomal drug resistance-conferring mutations. Antimicrob Agents Chemother. 2002;46(5):1204-11.
16. Gagneux S, Long CD, Small PM, Van T, Schoolnik GK, Bohannan BJ. The competitive cost of antibiotic resistance in *Mycobacterium tuberculosis*. Science. 2006;312(5782):1944-6.
17. Bottger EC, Springer B, Pletschette M, Sander P. Fitness of antibiotic-resistant microorganisms and compensatory mutations. Nat Med. 1998;4(12):1343-4.
18. Borrell S, Teo Y, Giardina F, Streicher EM, Klopper M, Feldmann J, et al. Epistasis between antibiotic resistance mutations drives the evolution of extensively drug-resistant tuberculosis. Evol Med Public Health. 2013;2013(1):65-74.
19. de Vos M, Muller B, Borrell S, Black PA, van Helden PD, Warren RM, et al. Putative compensatory mutations in the *rpoC* gene of rifampin-resistant *Mycobacterium tuberculosis* are associated with ongoing transmission. Antimicrob Agents Chemother. 2013;57(2):827-32.
20. Becerra MC, Huang C-C, Lecca L, Bayona J, Contreras C, Calderon R, et al. 2018.
21. Ronacher K, van Crevel R, Critchley JA, Bremer AA, Schlesinger LS, Kapur A, et al. Defining a Research Agenda to Address the Converging Epidemics of Tuberculosis and Diabetes. Chest. 2017;152(1):174-80.

22. Merker M, Barbier M, Cox H, Rasigade JP, Feuerriegel S, Kohl TA, et al. Compensatory evolution drives multidrug-resistant tuberculosis in Central Asia. Elife. 2018;7.
23. Comas I, Borrell S, Roetzer A, Rose G, Malla B, Kato-Maeda M, et al. Whole-genome sequencing of rifampicin-resistant *Mycobacterium tuberculosis* strains identifies compensatory mutations in RNA polymerase genes. Nat Genet. 2012;44(1):106-10.
24. Brandis G, Hughes D. Genetic characterization of compensatory evolution in strains carrying *rpoB* Ser531Leu, the rifampicin resistance mutation most frequently found in clinical isolates. J Antimicrob Chemother. 2013;68(11):2493-7.
25. Brandis G, Pietsch F, Alemayehu R, Hughes D. Comprehensive phenotypic characterization of rifampicin resistance mutations in Salmonella provides insight into the evolution of resistance in *Mycobacterium tuberculosis*. J Antimicrob Chemother. 2015;70(3):680-5.
26. Li QJ, Jiao WW, Yin QQ, Xu F, Li JQ, Sun L, et al. Compensatory Mutations of Rifampin Resistance Are Associated with Transmission of Multidrug-Resistant *Mycobacterium tuberculosis* Beijing Genotype Strains in China. Antimicrob Agents Chemother. 2016;60(5):2807-12.
27. Hubin EA, Fay A, Xu C, Bean JM, Saecker RM, Glickman MS, et al. Structure and function of the mycobacterial transcription initiation complex with the essential regulator RbpA. Elife. 2017;6.
28. Meyer MJ, Lapcevic R, Romero AE, Yoon M, Das J, Beltran JF, et al. mutation3D: Cancer Gene Prediction Through Atomic Clustering of Coding Variants in the Structural Proteome. Hum Mutat. 2016;37(5):447-56.
29. Cohen KA, Abeel T, Manson McGuire A, Desjardins CA, Munsamy V, Shea TP, et al. Evolution of Extensively Drug-Resistant Tuberculosis over Four Decades: Whole Genome Sequencing and Dating Analysis of *Mycobacterium tuberculosis* Isolates from KwaZulu-Natal. PLoS Med. 2015;12(9):e1001880.
30. Borrell S, Gagneux S. Strain diversity, epistasis and the evolution of drug resistance in *Mycobacterium tuberculosis*. Clin Microbiol Infect. 2011;17(6):815-20.
31. Chen ML, Doddi A, Royer J, Freschi L, Schito M, Ezewudo M, et al. Deep learning predicts tuberculosis drug resistance status from genome sequencing data. 2018.
32. Kouchaki S, Yang Y, Walker TM, Walker AS, Wilson DJ, Peto TEA, et al. Application of machine learning techniques to tuberculosis drug resistance analysis. Bioinformatics. 2018.
33. Comas I, Gagneux S. The past and future of tuberculosis research. PLoS Pathog. 2009;5(10):e1000600.
34. Brites D, Gagneux S. The Nature and Evolution of Genomic Diversity in the *Mycobacterium tuberculosis* Complex. Adv Exp Med Biol. 2017;1019:1-26.
35. Homolka S, Niemann S, Russell DG, Rohde KH. Functional genetic diversity among *Mycobacterium tuberculosis* complex clinical isolates: delineation of conserved core and lineage-specific transcriptomes during intracellular survival. PLoS Pathog. 2010;6(7):e1000988.
36. Rose G, Cortes T, Comas I, Coscolla M, Gagneux S, Young DB. Mapping of genotype-phenotype diversity among clinical isolates of *mycobacterium tuberculosis* by sequence-based transcriptional profiling. Genome Biol Evol. 2013;5(10):1849-62.
37. Zhu L, Zhong J, Jia X, Liu G, Kang Y, Dong M, et al. Precision methylome characterization of *Mycobacterium tuberculosis* complex (MTBC) using PacBio single-molecule real-time (SMRT) technology. Nucleic Acids Res. 2016;44(2):730-43.
38. Gomez-Gonzalez PJ, Andreu N, Phelan JE, de Sessions PF, Glynn JR, Crampin AC, et al. An integrated whole genome analysis of *Mycobacterium tuberculosis* reveals insights into relationship between its genome, transcriptome and methylome. Sci Rep. 2019;9(1):5204.
39. Fenner L, Egger M, Bodmer T, Altpeter E, Zwahlen M, Jaton K, et al. Effect of mutation and genetic background on drug resistance in *Mycobacterium tuberculosis*. Antimicrob Agents Chemother. 2012;56(6):3047-53.
40. Portevin D, Sukumar S, Coscolla M, Shui G, Li B, Guan XL, et al. Lipidomics and genomics of *Mycobacterium tuberculosis* reveal lineage-specific trends in mycolic acid biosynthesis. Microbiologyopen. 2014;3(6): 823-35.
41. Krishnan N, Malaga W, Constant P, Caws M, Tran TH, Salmons J, et al. *Mycobacterium tuberculosis* lineage influences innate immune response and virulence and is associated with distinct cell envelope lipid profiles. PLoS One. 2011;6(9):e23870.

42. Coscolla M. Biological and Epidemiological Consequences of MTBC Diversity. Adv Exp Med Biol. 2017;1019:95-116.
43. Stavrum R, PrayGod G, Range N, Faurholt-Jepsen D, Jeremiah K, Faurholt-Jepsen M, et al. Increased level of acute phase reactants in patients infected with modern *Mycobacterium tuberculosis* genotypes in Mwanza, Tanzania. BMC Infect Dis. 2014;14:309.
44. de Jong BC, Hill PC, Aiken A, Jeffries DJ, Onipede A, Small PM, et al. Clinical presentation and outcome of tuberculosis patients infected by *M. africanum* versus *M. tuberculosis*. Int J Tuberc Lung Dis. 2007;11(4):450-6.
45. Portevin D, Gagneux S, Comas I, Young D. Human macrophage responses to clinical isolates from the *Mycobacterium tuberculosis* complex discriminate between ancient and modern lineages. PLoS Pathog. 2011;7(3):e1001307.
46. Saelens JW, Viswanathan G, Tobin DM. Mycobacterial Evolution Intersects With Host Tolerance. Front Immunol. 2019;10:528.
47. Faksri K, Drobniewski F, Nikolayevskyy V, Brown T, Prammananan T, Palittapongarnpim P, et al. Epidemiological trends and clinical comparisons of *Mycobacterium tuberculosis* lineages in Thai TB meningitis. Tuberculosis (Edinb). 2011;91(6):594-600.
48. Caws M, Thwaites G, Dunstan S, Hawn TR, Lan NT, Thuong NT, et al. The influence of host and bacterial genotype on the development of disseminated disease with *Mycobacterium tuberculosis*. PLoS Pathog. 2008;4(3):e1000034.
49. Sinsimer D, Huet G, Manca C, Tsenova L, Koo MS, Kurepina N, et al. The phenolic glycolipid of *Mycobacterium tuberculosis* differentially modulates the early host cytokine response but does not in itself confer hypervirulence. Infect Immun. 2008;76(7):3027-36.
50. Palanisamy GS, DuTeau N, Eisenach KD, Cave DM, Theus SA, Kreiswirth BN, et al. Clinical strains of *Mycobacterium tuberculosis* display a wide range of virulence in guinea pigs. Tuberculosis (Edinb). 2009;89(3):203-9.
51. Merker M, Blin C, Mona S, Duforet-Frebourg N, Lecher S, Willery E, et al. Evolutionary history and global spread of the *Mycobacterium tuberculosis* Beijing lineage. Nat Genet. 2015;47(3):242-9.
52. Kato-Maeda M, Shanley CA, Ackart D, Jarlsberg LG, Shang S, Obregon-Henao A, et al. Beijing sublineages of *Mycobacterium tuberculosis* differ in pathogenicity in the guinea pig. Clin Vaccine Immunol. 2012;19(8):1227-37.
53. Ribeiro SC, Gomes LL, Amaral EP, Andrade MR, Almeida FM, Rezende AL, et al. *Mycobacterium tuberculosis* strains of the modern sublineage of the Beijing family are more likely to display increased virulence than strains of the ancient sublineage. J Clin Microbiol. 2014;52(7):2615-24.
54. Ates LS, Dippenaar A, Ummels R, Piersma SR, van der Woude AD, van der Kuij K, et al. Mutations in *ppe38* block PE_PGRS secretion and increase virulence of *Mycobacterium tuberculosis*. Nat Microbiol. 2018;3(2):181-8.
55. Gagneux S. Host-pathogen coevolution in human tuberculosis. Philos Trans R Soc Lond B Biol Sci. 2012;367(1590):850-9.
56. Be NA, Lamichhane G, Grosset J, Tyagi S, Cheng QJ, Kim KS, et al. Murine model to study the invasion and survival of *Mycobacterium tuberculosis* in the central nervous system. J Infect Dis. 2008;198(10):1520-8.
57. Jagielski T, Minias A, van Ingen J, Rastogi N, Brzostek A, Zaczek A, et al. Methodological and Clinical Aspects of the Molecular Epidemiology of *Mycobacterium tuberculosis* and Other Mycobacteria. Clin Microbiol Rev. 2016;29(2):239-90.
58. Faksri K, Xia E, Ong RT, Tan JH, Nonghanphithak D, Makhao N, et al. Comparative whole-genome sequence analysis of *Mycobacterium tuberculosis* isolated from tuberculous meningitis and pulmonary tuberculosis patients. Sci Rep. 2018;8(1):4910.
59. Falush D, Bowden R. Genome-wide association mapping in bacteria? Trends Microbiol. 2006;14(8):353-5.
60. Didelot X, Lawson D, Darling A, Falush D. Inference of homologous recombination in bacteria using whole-genome sequences. Genetics. 2010;186(4):1435-49.
61. Earle SG, Wu CH, Charlesworth J, Stoesser N, Gordon NC, Walker TM, et al. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. Nat Microbiol. 2016;1:16041.

62. Oppong YEA, Phelan J, Perdigão J, Machado D, Miranda A, Portugal I, et al. Genome-wide analysis of *Mycobacterium tuberculosis* polymorphisms reveals lineage-specific associations with drug resistance. BMC Genomics. 2019;20(1).
63. MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, et al. Guidelines for investigating causality of sequence variants in human disease. Nature. 2014;508(7497):469-76.
64. Nebenzahl-Guimaraes H, van Laarhoven A, Farhat MR, Koeken VA, Mandemakers JJ, Zomer A, et al. Transmissible *Mycobacterium tuberculosis* Strains Share Genetic Markers and Immune Phenotypes. Am J Respir Crit Care Med. 2016.
65. Lamichhane G, Bishai W. Defining the 'survivasome' of *Mycobacterium tuberculosis*. Nat Med. 2007;13(3): 280-2.
66. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature. 1998;393(6685):537-44.
67. Guerra-Assuncao JA, Crampin AC, Houben RM, Mzembe T, Mallard K, Coll F, et al. Large-scale whole genome sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area. Elife. 2015;4.
68. Casali N, Nikolayevskyy V, Balabanova Y, Harris SR, Ignatyeva O, Kontsevaya I, et al. Evolution and transmission of drug-resistant tuberculosis in a Russian population. Nat Genet. 2014;46(3):279-86.
69. Meehan CJ, Goig GA, Kohl TA, Verboven L, Dippenaar A, Ezewudo M, et al. Whole genome sequencing of *Mycobacterium tuberculosis*: current standards and open issues. Nat Rev Microbiol. 2019.
70. van Beek J, Haanpera M, Smit PW, Mentula S, Soini H. Evaluation of whole genome sequencing and software tools for drug susceptibility testing of *Mycobacterium tuberculosis*. Clin Microbiol Infect. 2019;25(1):82-6.
71. Phelan J, de Sessions PF, Tientcheu L, Perdigao J, Machado D, Hasan R, et al. Methylation in *Mycobacterium tuberculosis* is lineage specific with associated mutations present globally. Sci Rep. 2018;8(1):160.
72. Coll F, McNerney R, Preston MD, Guerra-Assuncao JA, Warry A, Hill-Cawthorne G, et al. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Med. 2015;7(1):51.
73. Chaidir L, Sengstake S, de Beer J, Oktavian A, Krismawati H, Muhapril E, et al. Predominance of modern *Mycobacterium tuberculosis* strains and active transmission of Beijing sublineage in Jayapura, Indonesia Papua. Infect Genet Evol. 2016;39:187-93.
74. Schleusener V, Koser CU, Beckert P, Niemann S, Feuerriegel S. *Mycobacterium tuberculosis* resistance prediction and lineage classification from genome sequencing: comparison of automated analysis tools. Sci Rep. 2017;7:46327.
75. Tagliani E, Cirillo DM, Ködmön C, van der Werf MJ, Anthony R, van Soolingen D, et al. EUSeqMyTB to set standards and build capacity for whole genome sequencing for tuberculosis in the EU. The Lancet Infectious Diseases. 2018;18(4):377.
76. Doyle RM, Burgess C, Williams R, Gorton R, Booth H, Brown J, et al. Direct Whole-Genome Sequencing of Sputum Accurately Identifies Drug-Resistant *Mycobacterium tuberculosis* Faster than MGIT Culture Sequencing. J Clin Microbiol. 2018;56(8).
77. Read TD, Massey RC. Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology. Genome Med. 2014;6(11):109.
78. Collins C, Didelot X. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. PLoS Comput Biol. 2018;14(2):e1005958.
79. Bartha I, Carlson JM, Brumme CJ, McLaren PJ, Brumme ZL, John M, et al. A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control. Elife. 2013; 2:e01123.
80. Kwong JC, McCallum N, Sintchenko V, Howden BP. Whole genome sequencing in clinical and public health microbiology. Pathology. 2015;47(3):199-210.
81. Satta G, Atzeni A, McHugh TD. *Mycobacterium tuberculosis* and whole genome sequencing: a practical guide and online tools available for the clinical microbiologist. Clin Microbiol Infect. 2017;23(2):69-72.

82. Iwai H, Kato-Miyazawa M, Kirikae T, Miyoshi-Akiyama T. CASTB (the comprehensive analysis server for the *Mycobacterium tuberculosis* complex): A publicly accessible web server for epidemiological analyses, drug-resistance prediction and phylogenetic comparison of clinical isolates. Tuberculosis (Edinb). 2015;95(6):843-4.
83. Steiner A, Stucki D, Coscolla M, Borrell S, Gagneux S. KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes. BMC Genomics. 2014;15:881.
84. Bradley P, Gordon NC, Walker TM, Dunn L, Heys S, Huang B, et al. Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. Nat Commun. 2015;6:10063.
85. Feuerriegel S, Schleusener V, Beckert P, Kohl TA, Miotto P, Cirillo DM, et al. PhyResSE: a Web Tool Delineating *Mycobacterium tuberculosis* Antibiotic Resistance and Lineage from Whole-Genome Sequencing Data. J Clin Microbiol. 2015;53(6):1908-14.
86. Sekizuka T, Yamashita A, Murase Y, Iwamoto T, Mitarai S, Kato S, et al. TGS-TB: Total Genotyping Solution for *Mycobacterium tuberculosis* Using Short-Read Whole-Genome Sequencing. PLoS One. 2015;10(11): e0142951.
87. Jajou R, van der Laan T, de Zwaan R, Kamst M, Mulder A, de Neeling A, et al. WGS more accurately predicts susceptibility of *Mycobacterium tuberculosis* to first-line drugs than phenotypic testing. J Antimicrob Chemother. 2019.
88. Starks AM, Aviles E, Cirillo DM, Denkinger CM, Dolinger DL, Emerson C, et al. Collaborative Effort for a Centralized Worldwide Tuberculosis Relational Sequencing Data Platform. Clin Infect Dis. 2015;61Suppl 3:S141-6.
89. [Available from: https://hivdb.stanford.edu.
90. Cowman S, van Ingen J, Griffith D, Loebinger MR. Non-tuberculous mycobacterial pulmonary disease. European Respiratory Journal. 2019:1900250.
91. van Ingen J, Kuijper EJ. Drug susceptibility testing of nontuberculous mycobacteria. Future Microbiol. 2014;9(9):1095-110.
92. Hirsh AE, Tsolaki AG, DeRiemer K, Feldman MW, Small PM. Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. Proc Natl Acad Sci U S A. 2004;101(14): 4871-6.
93. Tsolaki AG, Hirsh AE, DeRiemer K, Enciso JA, Wong MZ, Hannan M, et al. Functional and evolutionary genomics of *Mycobacterium tuberculosis*: insights from genomic deletions in 100 strains. Proc Natl Acad Sci U S A. 2004;101(14):4865-70.
94. Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, Narayanan S, et al. Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. Proc Natl Acad Sci U S A. 2006;103(8):2869-73.
95. van Crevel R, Parwati I, Sahiratmadja E, Marzuki S, Ottenhoff TH, Netea MG, et al. Infection with *Mycobacterium tuberculosis* Beijing genotype strains is associated with polymorphisms in SLC11A1/ NRAMP1 in Indonesian patients with tuberculosis. J Infect Dis. 2009;200(11):1671-4.
96. Lees JA, Ferwerda B, Kremer PHC, Wheeler NE, Seron MV, Croucher NJ, et al. Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis. Nat Commun. 2019;10(1):2176.
97. Yang Z, Zeng X, Tsui SK. Investigating function roles of hypothetical proteins encoded by the *Mycobacterium tuberculosis* H37Rv genome. BMC Genomics. 2019;20(1):394.
98. Ramakrishnan G, Ochoa-Montano B, Raghavender US, Mudgal R, Joshi AG, Chandra NR, et al. Enriching the annotation of *Mycobacterium tuberculosis* H37Rv proteome using remote homology detection approaches: insights into structure and function. Tuberculosis (Edinb). 2015;95(1):14-25.

8

Summary

Summary

The outcome of *M. tuberculosis* infection and the presentation and course of active tuberculosis disease are highly variable. Although many host and environmental factors have been identified that contribute to this variation, together they do not provide sufficient explanation for this variability. The growing knowledge of the global genomic diversity of *M. tuberculosis* complex (MTBC) affecting different cellular and immunological phenotypes supports that bacterial factors also play a role in the observed variation in clinical phenotypes of *M. tuberculosis* infection and active tuberculosis disease. Hence, in this thesis, containing whole genome sequencing data of more than 1,000 *M. tuberculosis* isolates and clinical metadata from tuberculosis patients from three different continents, I aimed to improve the understanding of bacterial factors determining drug resistance, disease presentation and transmission by studying genetic differences among *M. tuberculosis* strains on a genome-wide scale (summarized in **Table 8.1**).

Part one: drug resistance

The first part of this thesis focused on the use of whole genome sequencing to predict drug resistance in *M. tuberculosis* clinical isolates, and on the relation between diabetes and drug resistance mutations.

In **Chapter 2**, we sequenced a selection of 322 *M. tuberculosis* isolates from Indonesia to measure the distribution of drug resistance mutations and the concordance between phenotypic and genotypic drug susceptibility testing, which were until then unknown in Indonesia. We identified mutations associated with drug resistance to at least one tuberculosis drug in almost one sixth of the isolates. Most mutations were found in *katG*, *pncA*, *rpoB*, *fabG1* and *embB*. Agreement of whole genome sequence-based resistance prediction and phenotypic drug susceptibility testing to first-line drugs was high for isoniazid and rifampicin but was lower for ethambutol and streptomycin. Our findings support the potential benefit of using whole genome sequencing to generate an *in silico* drug susceptibility profile in Indonesia and show that mutations associated with drug resistance are highly predictive for phenotypic resistance to rifampicin and isoniazid in the region.

Whole genome sequencing may be a promising alternative to phenotypic drug susceptibility testing. However, the degree to which the various drug resistance mutations impact drug susceptibility remains to be investigated. Therefore, in **Chapter 3**, we examined drug resistance mutations in 72 phenotypically drug-resistant isolates from Romania and linked these to their measured minimum inhibitory concentrations. We observed that *katG* S315T for isoniazid, and *rpoB* S450L for rifampicin were associated with

Table 8.1. Summary of the main findings and implications of this thesis

Findings	Implications for practice or research
Part one: drug resistance	
Agreement of drug resistance mutations and phenotypic DST was high for INH and RIF but low for EMB and STR (Chapter 2), because low DST quality and resolution for these drugs (Chapter 3)	Genotypic DST could offer a rapid and comprehensive diagnostic solution to phenotypic DST in Indonesia; larger studies are needed to confirm the clinical relevance of several uncommon mutations (Chapter 2)
Drug resistance mutations lead to varying levels of resistance; some may be overcome by increased dosing (Chapter 3)	Whole genome sequencing can aid in the timely diagnosis of *M. tuberculosis* drug resistance and guide clinical decision-making (Chapter 3)
Diabetes is associated with drug-resistant TB, especially among patients without prior TB treatment (Chapter 4)	TB patients with diabetes should be prioritized for DST in settings where this is not done for all patients (Chapter 4)
Part two: disease phenotype and transmission	
Three genes in *M. tuberculosis* (*Rv0218*, *Rv3433c*, and *nanK*) are associated with the TB disease phenotype (Chapter 5)	Functional validation studies are warranted to further explore the effect of mutations in these genes on protein function (Chapter 5)
M. tuberculosis Beijing genotype strains may evade BCG vaccine-induced immunity (Chapter 6)	The impact of BCG vaccination on transmission may depend on the *M. tuberculosis* strain distribution in the setting where it is used (Chapter 6)

Abbreviations: DST: drug susceptibility testing; EMB: ethambutol; STR: streptomycin; TB: tuberculosis; BCG: Bacillus Calmette-Guérin.

high-level resistance (8- to 32-fold and 4- to 16-fold increase in MIC, respectively), but several mutations such as in *rpoB*, *rrs* and *rpsL*, and *embB* were associated with minimum inhibitory concentration ranges including the critical concentration for rifampicin, aminoglycosides and ethambutol, respectively. This could explain the discrepant genotypic and phenotypic susceptibility results for streptomycin and ethambutol in **Chapter 2**, and taught us that resistance mutations induce varying levels of resistance, some of which may still be overcome by increased dosing. Hence, whole genome sequencing can guide clinical decision-making, by informing drug selection as well as drug dosing.

Diabetes increases the risk of developing active tuberculosis by threefold and the increasing prevalence of diabetes poses a great threat to tuberculosis control. Increased rates of drug resistance may contribute to poor treatment outcomes among tuberculosis patients with diabetes, complicating this even further. Still, little is known about the relation of diabetes and genotypic drug resistance. Therefore, in **Chapter 4**, we investigated the mutations underlying drug resistance in *M. tuberculosis* isolates from Indonesian and Peruvian tuberculosis patients with and without diabetes. Taking patients' country of origin into account and after adjusting for age, gender, previous tuberculosis treatment and *M. tuberculosis* lineage, we discovered that diabetes was associated with an approximately 1.7-fold increased risk of genotypic drug resistance, and to rifampicin in particular (2.5-fold increased risk). This association was especially evident for patients not previously treated for tuberculosis ('primary resistance'), and could not be explained by clustering of evolutionarily successful drug-resistant strains in patients with diabetes. Mutations in *Rv1482c-fabG1*, conferring resistance to isoniazid and ethionamide, and *gyrA*, conferring fluoroquinolone resistance, were overrepresented in isolates from Peruvian tuberculosis patients with diabetes. These results are important for the control of drug-resistant tuberculosis as they highlight the need for routine drug-susceptibility testing for tuberculosis patients with diabetes.

Part two: tuberculosis disease phenotype and transmission

The second part of this thesis focused on the role of *M. tuberculosis* genetic diversity in the presentation of disease as pulmonary or meningeal tuberculosis, and on disease transmission.

To determine whether *M. tuberculosis* genetic diversity is associated with the tuberculosis disease phenotype, we compared *M. tuberculosis* genomes from Indonesian pulmonary and meningeal tuberculosis patients and searched for homoplastic mutations (**Chapter 5**). We observed that the isolates from meningitis patients were scattered across the phylogenetic tree and that there was no association between lineage and disease localisation. At a higher level of resolution and using a homoplasy-based association analysis, we found that genetic variation in *Rv0218* and absence of *Rv3343c* and *nanK* were significantly associated with disease phenotype, also in an independent validation set of isolates. It is unknown how *Rv3343c* and *nanK* might affect tuberculosis phenotype, but *Rv0218* is a predicted secretome gene with multiple transmembrane regions, possibly altering *M. tuberculosis* outer structure.

Similar to *M. tuberculosis* tissue tropism and supported by previous findings that *M. tuberculosis* virulence is associated with the isolate's genetic background, the risk of *M. tuberculosis* transmission could also be influenced by its genotype. Hence, we studied the effect of *M. tuberculosis* lineage on the risk of transmission from index

cases to household contacts in Bandung, Indonesia (**Chapter 6**). We discovered that *M. tuberculosis* Beijing genotype strains are associated with a higher risk of uninfected household contacts to acquire *M. tuberculosis* infection (measured with an interferon-gamma release assay). This was not explained by a higher bacillary load or higher frequency of pulmonary cavities in the index cases. In addition, a history of BCG vaccination among household contacts was associated with a 60% lower risk of *M. tuberculosis* infection caused by non-Beijing isolates, but no lower risk of infection by Beijing family isolates. Evading BCG-mediated protection in the contact, rather than inducing a disease phenotype in the index case that favours transmission, may increase the transmission rate in Beijing genotype strains.

9

Samenvatting

Samenvatting

Mycobacterium tuberculosis

Tuberculose is met 1,7 miljoen sterfgevallen in 2017 een van de top-10 doodsoorzaken ter wereld en de nummer één doodsoorzaak als gevolg van een infectieziekte. Naar schatting tien miljoen mensen wereldwijd kregen de ziekte in 2017. De ziekteverwekker, *Mycobacterium tuberculosis* behoort tot een grotere famile van aan elkaar verwante mycobacteriën: het *Mycobacterium tuberculosis* complex; het DNA van deze bacteriën lijkt sterk op elkaar. De bacteriën die tuberculose in de mens veroorzaken, kunnen weer onderverdeeld worden in 7 verschillende genotypen met elk specifieke genetische kenmerken. Lang dacht men dat er zo weinig genetische variatie binnen de tuberculose-bacteriën was, dat dit onmogelijk effect kon hebben op de manier waarop de ziekte zich uit in de patiënt. Echter, mede door vooruitgang in de techniek waarmee het DNA bestudeerd kan worden, is aangetoond dat verschillen in het DNA van de bacterie wel degelijk van invloed zijn op het ziekteproces. In dit proefschrift heb ik door middel van '*whole genome sequencing*' het complete DNA van *M. tuberculosis* onder de loep genomen om het effect van genetische variatie op antibioticaresistentie (deel 1), ziekte-fenotype (deel 2) en transmissie van tuberculose (deel 2) beter te begrijpen. In dit hoofdstuk vat ik de belangrijkste bevindingen van mijn proefschrift samen.

Deel één: antibioticaresistentie

Een van de grootste zorgen met betrekking tot tuberculosebestrijding is resistentie van de bacterie tegen een of meerdere geneesmiddelen. Hierdoor is de infectie minder goed of in zeldzame gevallen helemaal niet meer te behandelen. Slechts een kwart van de naar schatting 558,000 mensen die ziek werden van een multiresistente of rifampicine-resistente bacterie in 2017 werden met tweedelijnsmiddelen behandeld. Een belangrijke oorzaak hiervan is het niet, of niet op tijd diagnosticeren van infectie met een resistente bacterie. Wanneer ongediagnisticeerde resistente tuberculose niet volgens de juiste richtlijn wordt behandeld zal deze niet genezen en kan de resistente bacterie bovendien worden overgedragen op andere ongeïnfecteerde personen. Omdat antibioticaresistentie in *M. tuberculosis* veroorzaakt wordt door mutaties in bepaalde genen van de bacterie kan resistente tuberculose worden opgespoord door te zoeken naar deze mutaties in het DNA van de infecterende bacterie: genotypische gevoeligheidsbepaling. Veelal wordt dit echter nog gedaan door middel van de zogenaamde fenotypische gevoeligheidbepaling, waarbij in het laboratorium wordt vastgesteld of de bacterie groeit bij een bepaalde concentratie antibioticum.

De fenotypische gevoeligheidsbepaling is complex en bemoeilijkt de diagnose van resistente tuberculose. Whole genome sequencing heeft de potentie om betrouwbaar antibioticaresistentie te kunnen voorspellen binnen een klinisch relevant tijdsbestek. In **hoofdstuk 2** hebben we met whole genome sequencing de mutaties in kaart gebracht die voor resistentie zorgen bij *M. tuberculosis* stammen van Indonesische patiënten. Daarnaast hebben we onderzocht in hoeverre de genotypische en fenotypische gevoeligheidsbepalingen overeen kwamen. We ontdekten bij bijna één op de zes *M. tuberculosis* stammen mutaties die geassocieerd zijn met antibioticaresistentie, vaak bij patiënten die niet eerder behandeld waren voor tuberculose. We vonden veel genetische varianten in verschillende genen. Mutaties in de genen *katG* (aminozuur 315), in de promotor van *fabG1* en in de bekende resistentieregio van rpoB lieten hoge overeenstemming zien met fenotypisch vastgestelde resistentie tegen respectievelijk isoniazide en rifampicine. Deze overeenkomst was veel lager voor mutaties geassocieerd met resistentie tegen ethambutol en streptomycine. We hebben hiermee voor het eerst de potentie van whole genome sequencing als voorspeller van resistente tuberculose laten zien in Indonesië, waar het fenotypisch bepalen van resistentie vaak niet mogelijk is. Met name mutaties geassocieerd met resistentie tegen de eerstelijns middelen isoniazide en rifampicine waren zeer voorspellend voor fenotypische resistentie in deze setting. Verder onderzoek is nodig om de impact van whole genome sequencing voor routine diagnostiek en management van resistente tuberculose in Indonesië te kunnen evalueren.

Whole genome sequencing kan in potentie de huidige, maar tijdrovende en complexe fenotypische gevoeligheidsbepaling van *M. tuberculosis* vervangen. Echter, er is nog weinig bekend over de mate van resistentie die de verschillende resistentiemutaties veroorzaken en over het effect van combinaties van resistemntiemutaties. Daarom hebben we in **hoofdstuk 3** de met antibioticaresistentie geassocieerde mutaties in fenotypisch resistente *M. tuberculosis* stammen van Roemeense patiënten vergeleken met de mate van verminderde gevoeligheid tegen de verschillende antibiotica. We ontdekten dat de mutaties *katG* S315T en *rpoB* S450L geassocieerd waren met sterk verminderde gevoeligheid voor respectievelijk isoniazide en rifampicine en dat diverse mutaties in *rpoB*, *rrs* en *rpsl*, en *embB* geassocieerd waren met licht verminderde gevoeligheid tegen respectievelijk rifampicine, aminoglycosiden en ethambutol. Hierbij werden ook waarden gevonden rond het omslagpunt van gevoelig / resistent; dit zou mogelijk de lage overeenstemming tussen de fenotypische en genotypische gevoeligheidsbepaling voor ethambutol en streptomycine in **hoofdstuk 2** kunnen verklaren. Verschillende resistentiemutaties kunnen dus verschillende effecten hebben op de mate van gevoeligheid van *M. tuberculosis* voor antibiotica en in sommige gevallen kan de patiënt mogelijk nog effectief behandeld worden met een hogere dosis antibiotica. Whole genome sequencing kan hiermee richting geven aan de klinische besluitvorming met betrekking tot de keuze voor het juiste middel en de hoogte van de dosering.

Mensen met diabetes mellitus hebben een drie keer verhoogd risisco op het ontwikkelen van actieve tuberculose. Het toenemende aantal diabetespatiënten wereldwijd heeft dus gevolgen voor de tuberculose-epidemie. Een groter aandeel resistente tuberculose onder diabetespatiënten vormt een verdere bedreiging voor de tuberculosebestrijding. Toch is er nog weinig bekend over de resistentiemutaties in deze speciefieke patiëntenpopulatie. In **hoofdstuk 4** hebben we daarom de resistentiemutaties bestudeerd in *M. tuberculosis* stammen van Indonesische en Peruaanse tuberculosepatiënten met en zonder diabetes. We vonden dat diabetes geassocieerd was met een 1.7 keer verhoogd risisco op genotypisch resistente tuberculose en met name resistentie tegen rifampicine. Dit kon niet verklaard worden door andere patiëntegrelelateerde factoren zoals leeftijd, geslacht en eerdere behandeling voor tuberculose, of het genotype waartoe de *M. tuberculosis* stam behoorde. Bovendien zagen we geen verband met transmissie van evolutionair gezien meer succesvolle resistente stammen in diabetes patiënten. Mutaties in met name de genen *Rv1482-fabG1*, geassocieerd met resistentie tegen isoniazide en ethionamide, en *gyrA*, geassocieerd met fluoroquinolonenresistentie, kwamen vaker voor onder Peruaanse tuberculosepatiënten met diabetes. Deze resultaten zijn belangrijk omdat ze het belang van de invoering van een routine gevoeligheidsbepaling van *M. tuberculosis* in deze patiëntengroep aantonen.

Deel twee: ziektefenotypen en transmissie van tuberculose

M. tuberculosis wordt vooral verspreid via de lucht wanneer een tuberculosepatiënt hoest of niest. Wanneer iemand anders deze bacterie vervolgens inademt, zijn er verschillende scenario's mogelijk: een deel van de blootgestelde individuen ruimt *M. tuberculosis* op, nog voordat het afweersysteem hiertegen geheugen opbouwt. Een kleine minderheid wordt direct en vaak ernstig ziek, maar de meerderheid merkt niets van de besmetting. De bacterie verkeert dan in een 'slapende' staat en de patiënt is niet besmettelijk. De patiënt kan later in het leven alsnog ziek worden. De tuberculosebacterie kan namelijk tientallen jaren in slapende staat in het lichaam overleven en weer actief worden als de afweer van de geïnfecteerde persoon afneemt. Dat gebeurt bij omgeveer één op de tien besmette personen. Actieve tuberculose manifesteert zich doorgaans in de long, maar *M. tuberculosis* kan zich in alle organen nestelen en daar ziekte veroorzaken. Tuberculose in de hersenen, tuberculeuze meningitis, is de meest ernstige vorm van tuberculose. De helft van de patiënten houdt er zware neurologische handicaps aan over of overlijdt aan de aandoening. Uitzonderingen daargelaten, kunnen alleen patiënten met actieve longtuberculose de bacterie overdragen op anderen.

Verschillende factoren bepalen welk scenario plaatsvindt na blootstelling aan *M. tuberculosis* en hoe de ziekte zich eventueel uit in de patiënt. Veel van deze factoren zijn patiënt- of omgevingsgerelateerd. Echter, we komen er steeds meer achter dat ook de

bacterie een belangrijke rol speelt. De mate waarin bacteriën 'ziekmakend' (virulent) zijn verschilt, en genetische variatie binnen *M. tuberculosis* kan aan deze verschillen ten grondslag liggen. In **hoofdstuk 5** hebben we bekeken of genetische verschillen tussen *M. tuberculosis* bacteriën geassocieerd zijn met het meningitis of long-tuberculose fenotype door het DNA van bacteriën uit de hersenen van tuberculeuze meningitispatiënten te vergelijken met het DNA van bacteriën uit de longen van long-tuberculosepatiënten. We vonden dat het *M. tuberculosis* genotype niet bepalend was, maar dat subtielere genetische verschillen geassocieerd waren met het ziektefenotype. Mutaties in het gen *Rv0218* en de afwezigheid van gen *Rv3343c* of *nanK* hielden verband met uiting van de ziekte in de longen of in de hersenen. Gebaseerd op biologische eigenschappen van het gen *Rv0218* zou men kunnen veronderstellen dat dit gen invloed heeft op de buitenkant - en daarmee op de herkenning door het afweersysteem - van de tuberculosebacterie. Meer onderzoek is nodig naar de manier waarop variatie in deze genen invloed heeft op het ziektefenotype.

Naast het ziektefenotype kan genetische variatie binnen *M. tuberculosis* ook invloed hebben op de mate van overdraagbaarheid van tuberculose. Dit kan bijvoorbeeld doordat een bacterie meer schade in de longen aanricht waardoor de patiënt meer hoest, of doordat de bacterie kan ontsnappen aan het afweersysteem van een blootgesteld individu. In **hoofdstuk 6** onderzochten we de overdraagbaarheid van *M. tuberculosis* op mensen binnen hetzelfde huishouden in Bandung, Indonesië. We kwamen erachter dat ongeïnfecteerde huisgenoten die blootgetseld waren aan een patiënt besmet met een *M. tuberculosis* stam van het *Beijing* genotype, een hoger risico hadden om besmet te raken. Dit kon niet verklaard worden door een hogere concentratie tuberculosebacteriën in, of meer schade aan de longen van de patiënt. Bovendien ontdekten we dat huisgenoten die een BCG-vaccinatie hadden gehad, sterk beschermd waren tegen besmetting na blootstelling aan *M. tuberculosis,* maar dat BCG-vaccinatie helemaal niet beschermde tegen besmetting bij huisgenoten die blootgesteld waren aan *M. tuberculosis* genotype Beijing. Kortom, het ontsnappen aan BCG-gemediëerde bescherming lijkt de verhoogde overdraagbaarheid van *M. tuberculosis* genotype Beijing stammen te verklaren.

Al met al biedt het analyseren van het complete DNA van de tuberculosebacterie mogelijkheden, zowel voor de diagnostiek van resistente tuberculose als voor het begrijpen van deze complexe ziekte, om haar uiteindelijk beter te kunnen bestrijden. Dit vergt het blijvend bijhouden van de correlaties tussen het genotype van de bacterie en de verschillende fenotypische aspecten, zoals ik in mijn proefschrift heb gedaan. Ik hoop dat de bevindingen in mijn proefschrift bijdragen aan de ontwikkeling van betere hulpmiddelen en strategieën om tuberculose te kunnen beheersen. Een van de grootste uitdagingen zal liggen in het toegankelijk maken van deze technieken in

landen met weinig middelen, waar de ziektelast het hoogst is en de middelen het hardst nodig zijn.

Appendix

Dankwoord

Mijn dank gaat uit naar iedereen die mij tijdens mijn promotie op welke manier dan ook heeft bijgestaan, gesteund, geïnspireerd, geholpen en gemotiveerd. Bedankt dat jullie me hebben vergezeld op deze wetenschappelijke ontdekkingsreis.

In het bijzonder wil ik mijn promotores en copromotores bedanken. Ik voel me gezegend met zo een divers gezelschap inspirerende mensen om me heen. Reinout van Crevel, je hebt me alle vrijheid en vertrouwen gegeven om de wereld van de tuberculose en de onderzoekswereld in bredere zin te verkennen. Met je aanstekelijke enthousiasme heb je me uitgedaagd en de mogelijkheden gegeven om letterlijk en figuurlijk mijn grenzen te verleggen. Dankjewel voor alles wat je me hebt gegeven, maar zeker ook voor alles wat je van me hebt gevraagd. Martijn Huynen, je hebt me geïntroduceerd en warm welkom geheten in de wondere wereld van de bioinformatica. Je eindeloze nieuwsgierigheid en drang om alles te willen begrijpen werkt inspirerend. Voor jou is ieder antwoord slechts het begin van een nieuwe vraag. Jakko van Ingen, ik ben je bijzonder dankbaar voor de mogelijkheid die je me hebt geboden om mijn promotieonderzoek af te ronden. Met je bezielende positiviteit en enthousiasme creëer je een prettige sfeer en neem je anderen mee in je fascinatie voor de niet-tuberculeuze mycobacteriën. Lidya Chaidir, dear Lidya, I admire your dedication to tuberculosis research and your capability to combine so many different tasks; doing a PhD, teaching, managing the tuberculosis lab in Bandung, supervising and travelling. You have become a dear friend and I am proud to have you as my copromoter. Dank allen voor jullie vertrouwen in mij!

De leden van de manuscriptcommissie, Michel van den Heuvel, Wilbert Bitter en Jeannine Hautvast wil ik bedanken voor hun bereidheid om zitting te nemen in de manuscriptcommissie en voor het lezen en beoordelen van dit proefschrift.

Ik wil ook graag de mensen bedanken die mijn mentor en voorbeeld zijn geweest tijdens eerdere onderzoeksprojecten. Ann Ashburn, Vivian Weerdesteyn en Geert Verheyden namen me onder hun vleugels tijdens en voorafgaand aan mijn project in Southampton. Rob Aarnoutse bood me samen met Reinout van Crevel de kans om een dataset te analyseren en hierover te publiceren. Dick van Soolingen en Rianne van Gageldonk hebben me wegwijs gemaakt in de epidemiologie van tuberculose. Dick, ik vind het leuk dat we elkaar af en toe nog tegenkomen en ik hoop dat we elkaar tegen blijven komen. Marlies Hulscher was mijn mentor zowel tijdens mijn afstudeerproject als tijdens mijn promotieonderzoek. Marlies, ik bewonder je gedrevenheid in het verbeteren van de kwaliteit van zorg en ik mag je erg graag als mens.

The research I did during my PhD was in close collaboration with many inspiring colleagues from Bandung, where I worked and lived for seven months, and from many other countries. I am grateful for the lessons I learned from you, not only about science, but also about hospitality, modesty, and gratefulness. Dr. Bachti Alisjahbana, thank you for kindly welcoming me as a member of the TB/HIV research group and for looking after me, also outside the office. Professor Rovina Ruslami, thank you for being such a welcoming host. Sue McAllister, dear Sue, thank you for generously sharing your house with me. I learned a lot from you about doing research and living in a foreign country. I cherish the memories of our weekend breakfasts outside and the many motorbike rides. Professor Philip Hill, thank you for showing that doing high quality research and a good laugh can go hand in hand. Ayesha Verrall, dear Ayesha, you've inspired me both professionally as well as personally. I enjoyed our epidemiology discussions. Anca Riza, dear Anca, your expertise is human genetics but you have been a great help with mycobacterial genetics analyses, not to forget the shipments of mycobacterial DNA. My gratitude also goes out to 'the ladies' from the tuberculosis lab in Craiova for all their hard work on culturing and managing isolates. I would like to thank Julia Critchley, Cesar Ugarte, David Moore and other colleagues from the TANDEM consortium for contributing to Chapter 4 of my thesis. Thomas Kohl and Matthias Merker, thank you for introducing me to whole genome sequencing of *M. tuberculosis*. Dr. Ahmad Rizal Ganiem and Dr. Sofiati Dian, you are doing important scientific and clinical work on tuberculous meningitis. I learned a lot from you, also about the Indonesian cuisine. Dr. Darma Imran, thank you for giving me the opportunity to visit the emergency ward at the Cipto Mangunkusumo hospital in Jakarta. I would also like to thank my other Indonesian colleagues, Raspati Koesoemadinata, Intan Mauli, Tiara Pramaesya, Annisa 'Itcha' Rahmalia, Mawar Pohan, Ria Windyani, Bony Wiem Lestari, Jessi Annisa, Lika Apriani, my former roommates Ranni Trisnawati, Runi Rahmawati, Nopi Susilawati and Alif al Birru, the colleagues from the tuberculosis lab in Bandung who took great care of the M. tuberculosis isolates, and all others who contributed to the work presented in this thesis but are not mentioned here explicitly.

The time I spent in Bandung was not only an academic learning experience, it was a period of personal development. I am grateful for all the people I met and the friends I made, also outside the office. Misliani Intan Ardita, dear 'Dita', my football buddy. I really enjoyed our afternoon practices with the Rumah Cemara team. Thanks for welcoming me to your football teams and for introducing me to so many Indonesian traditions. Itcha, thank you for taking me by the hand on my quest for a football team to join in Bandung. Rizki Kurniawan, the way you use your hobby to do meaningful work is inspirational. Our visit to Persib Bandung was memorable! Ferlin Joswara, your work spirit is amazing and you create the most wonderful things. It was an honour to model for you. I am also grateful for the chance I got to play football with the girls

from Rumah Cemara, run with the Indorunners and discover the Bandung mountains with the Bandung biking group. You have helped keeping mind and body healthy. Arjan van Laarhoven en Suzanne van Dorp, dank jullie wel voor jullie grote dosis gastvrijheid tijdens mijn tweede verblijf in Bandung. Met de warmte van jullie persoonlijkheden en de verfrissing van koude biertjes en het zwembad voelde jullie huis als een tweede thuis.

Een van de leuke dingen aan promoveren is de vrijheid om je helemaal in een onderwerp te mogen verdiepen in een omgeving van talentvolle mensen van (en met) wie je veel kan leren. Ik heb het geluk gehad dat ik tijdens mijn promotieonderzoek heb mogen werken op drie verschillende afdelingen binnen het Radboudumc; de interne geneeskunde, het center for molecular and biomolecular informatics (CMBI) en de medische microbiologie. Van collega's met verschillende achtergronden heb ik geleerd om vraagstukken vanuit verschillende invalshoeken te benaderen. Met veel plezier heb ik samengewerkt met andere (ex-) promovendi aan tuberculose en niet-tuberculeuze mycobacteriën. Arjan van Laarhoven, Lidya Chaidir, Sofiati Dian, Lindsey te Brake, Valerie Koeken, Edwin Ardiansyah, Ekta Lachmandas, Hinta Meijerink, Anca Riza, Hanna Guimaraes, Jodie Schildkraut, Mike Ruth en Jordy Coolen, ik heb ervan genoten om met jullie na te denken over ingewikkelde concepten, vragen, scripts en analyses en bovenal ook van de gezelligheid. Daarnaast gaat mijn dank uit naar ervaren wetenschappers binnen en buiten het Radboudumc van wie ik veel heb geleerd, Mihai Netea, Jos van der Meer, Heiman Wertheim, Vinod Kumar, Alexander Hoischen, Jelle Goeman, Jan Buitelaar, Taco Kooij, Hester Korthals-Altes, Aldert Zomer, Sacha van Hijum, Han de Neeling en Bas Dutilh, zonder wie ik nooit had leren programmeren. Mijn nieuwe kamergenoten bij de medische microbiologie, John Verbeek, Marion Dinnissen–van Poppel, Nel Gerrits-Hofmans, Paul Rutten, Jordy Coolen, Tim Baltussen, Margriet Hokken en Jan Zoll ben ik dankbaar voor het warme welkom, net als Ingrid van Weerdenburg, Saskia Kuipers, Ellen Koenraad, Mariëlle Rockland, Lian Pennings, Mike Mientjes, Melanie Wattenberg, Nicole Aalders, Jasper Sangen, Twan Klaassen en Arjan de Jong. Veel bioinformatica-inzichten heb ik gekregen dankzij (oud-) collega's bij het CMBI; Peter-Bram 't Hoen, Gert Vriend, John van Dam, Wynand Alkema, Rob ter Horst, Jon Black, Wouter Touw, Xiaowen Lu, Selma van Esveld, Barbara van Kampen, Josh Gillard, William Leenders, Daniel Garza, Balaji Venkatasubramanian, Laurens van de Wiel, Joeri van Strien, Tom Ederveen, Hanka Venselaar, Charlotte Kaffa, Sander Bervoets, Coos Baakman, Jeron Venhuizen, Renee Salz en niet te vergeten Arthur Pistorius, die menig serverprobleem voor me heeft opgelost. In particular, I'd like to thank Dei Elurbe, Joanna Lange, Robin van der Lee and Lisette Meerstein-Kessel. You are not only skilled bioinformaticians, but also close friends. I hope we will stay in touch, wherever the wind may take us. Dear Lise, I am verry happy to have you as my paranimph.

Ik wil ook graag mijn dank betuigen aan Barbara van Kampen, Mieke Daalderop en Jeanine van der Wijst-Schrauwen, die ontzettend veel werk uit handen nemen van artsen en onderzoekers in de dynamische academische ziekenhuisomgeving.

Voor een promovendus is de verleiding groot om je alleen te focussen op de inhoud van je onderzoek. Ik ben blij dat ik een groot deel van mijn promotietraject lid ben geweest van de PhD Council van het Radboud Institute for Health Sciences. Het heeft mijn blik verruimd en me bovendien de mogelijkheid gegeven om iets te kunnen betekenen voor medepromovendi. Ik ben Bart Kiemeney, Karin Berens, Marieke de Visser, Marie-Louise Roovers en Dagmar Eleveld-Trancikova dankbaar voor het vertrouwen en de verantwoordelijkheden die ze ons hebben gegeven. Alle andere leden en oud-leden van de Council wil ik graag bedanken voor jullie inzichten, het fijne samenwerken en de altijd gezellige sfeer. Ik hoop jullie nog vaak tegen te komen in de toekomst.

Jan en Doreth, dank jullie wel dat jullie deur altijd voor me open staat. Ik heb grote bewondering voor jullie positiviteit en levenslust.

Hard werken vraagt ook de nodige ontspanning. Mijn (oud-) teamgenootjes en trainster bij Trekvogels wil ik graag bedanken voor de energie die jullie me hebben gegeven door jullie enthousiasme en de drive om samen te willen winnen. Voetbal is voor mij de ideale uitlaatklep en met jullie heb ik een team gevonden dat prestatiedrang koppelt aan teamspirit en gezelligheid. Ik hoop met jullie een mooie herfst van mijn voetbalcarrière in te gaan. Ik ben dankbaar voor alle lieve vrienden om me heen met wie ik lief en leed kan delen en die me met een gezonde dosis humor helpen relativeren en ontspannen. Rozemarijn, Erika, Sander, Jilske, Mirjam, Ieke, Jorie en Sanne, ik ben blij met jullie in mijn leven.

Soms moet je de halve wereld over om iemand die dichtbij woont te leren kennen. Lieve Rachel, in Bandung ontmoette ik je en intussen ben je een dierbare vriendin geworden. Je beseft zelf denk ik maar half hoeveel je me hebt gesteund de afgelopen jaren. Dankjewel voor al je positiviteit, lieve attenties en je vermogen om altijd een lach op mijn gezicht te kunnen toveren.

En soms moet je iemand anders tegenkomen om jezelf beter te leren kennen. Lieve Stéphanie, dankjewel voor je open en verfrissende blik op het leven en voor onze fijne vriendschap die ik zeer waardeer.

Lieve Charlotte, Svenja en Marloes, ik vind het zo bijzonder dat wij alle vier aan een promotieonderzoek zijn begonnen. Jullie weten als geen ander hoe het is om te promoveren en hebben aan een half woord genoeg. Jullie staan altijd voor me klaar, of

het nou is om me op te beuren na tegenslag of om een feestje te vieren na goed nieuws. Laten we dat ook hierna zo blijven doen. Marloes, je hebt het goede voorbeeld gegeven en ik ben trots dat ik jou als paranimf naast me heb staan.

Promoveren betekent soms met oogkleppen op je afsluiten voor de wereld en diep de materie induiken. Het is dan heel fijn om daar af en toe even uitgehaald te worden door goede vrienden die niets met promoveren te maken hebben. Lieve Sophie, Veronique en Marije, de inhoud van mijn proefschrift zegt jullie helemaal niets ("Wat onderzoek je ook alweer? Parageen genomische tuberculose?"), maar mede dankzij jullie is deze nu wel af. Dank jullie wel voor jullie steun, positivisme, blijheid en gekkigheid. Ik ken jullie al meer dan mijn halve leven en ik zou nooit zonder jullie willen.

Ik ben blij en voel me gezegend met mijn lieve, warme familie en iedereen die ik als familie beschouw. In het bijzonder wil ik opa en oma Maandag bedanken. Lieve opa, je kan er helaas niet meer bij zijn, maar ik weet zeker dat je trots op me geweest zou zijn. Lieve oma, ik had me geen lievere oma kunnen wensen. Dankjewel voor jullie liefde en steun die ik altijd heb gevoeld en nog steeds voel.

Lieve Tineke, je bent als een moeder voor me en staat altijd voor me klaar. Je bent een voorbeeld door wie je bent en hoe je in het leven staat.

Lieve Suzan, mijn wetenschappelijke ontdekkingsreis zit er nu op, maar ons avontuur is pas net begonnen. Gelukkig is er nog een hele wereld te ontdekken. Ik kijk ernaar uit om samen met jou in het leven te verdwalen.

Ik ben trots op waar ik vandaan kom en dankbaar voor de normen en waarden die ik van huis uit heb meegekregen. Een warm en liefdevol thuis met lieve ouders en een zusje waar ik altijd op terug kan vallen is met geen titel te verkrijgen. Lieve Isabel, hoewel we enorm verschillen zijn we over de jaren heel close geworden. Ik ben je dankbaar voor je eindeloze geduld en de leuke momenten die we samen hebben beleefd. Ik vind het leuk hoe we elkaar aanvullen en van elkaar leren. De gedeelde liefde voor wijn heeft ons al op leuke plekken gebracht en ik ben benieuwd wat er nog meer in het vat zit. Lieve papa en mama, jullie geloof in mij is hartverwarmend en alle woorden schieten tekort om jullie te bedanken. Het begripvolle, empathische en bedachtzame van jou, papa, en het liefdevolle en zorgzame van jou, mama, draag ik altijd bij me. Jullie hebben me geleerd om op mijn gevoel te vertrouwen. Met humor en relativeringsvermogen hebben jullie me op vele momenten bijgestaan en het plezier en de positiviteit waarmee jullie in het leven staan is inspirerend. Bedankt dat jullie een omgeving hebben gecreëerd waarin ik me veilig en geliefd voel en kon groeien tot wie ik nu ben.

Curriculum Vitae

Carolien Ruesen was born in Wehl, the Netherlands, on July 26th in 1988, and grew up in Duiven with her parents and sister. She graduated from the Stedelijk Gymnasium Arnhem, after which she studied Biomedical Sciences at the Radboud University in Nijmegen. Carolien went to Southampton, the United Kingdom, for her Bachelor's internship at the Southampton General Hospital where she studied spinal posture in chronic stroke patients under supervision of professor Ann Ashburn. She also participated in the Honours programme for Medical Sciences. During her Master Biomedical Sciences she did a major in Epidemiology, and gained her first experience with tuberculosis research during an extracurricular research project on tuberculosis together with dr. Rob Aarnoutse and professor Reinout van Crevel. Then, she performed an internship on drug-resistant tuberculosis at the Rijksinstituut voor Volksgezonheid en Milieu with professor Dick van Soolingen and dr. Rianne van Gageldonk. Carolien performed her graduate internship – studying determinants of HIV viral 'blips' - at the Radboud university medical center department of IQ Healthcare in Nijmegen, under supervision of professor Marlies Hulscher and professor Reinout van Crevel, for which she was awarded the Radboudumc Masterprijs Biomedische wetenschappen, and the Best Student Award Epidemiologie from the department for Health Evidence. After graduation, Carolien participated in the Radboud university medical center PhD Competition and received a scholarship to start a PhD project designed by herself, under supervision of professor Van Crevel. In preparation for her PhD, she worked for the TB-HIV research group led by dr. Bachti Alisjahbana at the Padjadjaran University in Bandung, Indonesia, and lived there for seven months. Besides her research activities, Carolien has been a member of the Radboud Institute for Health Sciences PhD Council. She is currently working at the Radboudumc department of medical microbiology, in the mycobacteriology lab led by dr. Jakko van Ingen.

PhD portfolio

Institute for Health Sciences
Radboudumc

Name PhD candidate:
C.J. Ruesen, MSc.
Department:
Internal Medicine / Medical Microbiology
Graduate School:
Radboud Institute for Health Sciences

PhD period:
01-04-2014 – 01-07-2019
Promotor(s):
Prof. R. van Crevel, Prof. M.A. Huynen
Co-promotor(s): Dr L. Chaidir, Dr J. van Ingen

	YEAR(S)	ECTS
TRAINING ACTIVITIES		
A) COURSES & WORKSHOPS		
RIHS Introduction course for PhD students	2014	1.0
Introduction to R course, Maastricht University	2014	1.0
5E005 Genetic epidemiology	2014	5.5
Scientific Integrity	2015	1.0
How to write a medical scientific paper	2015	0.2
Comparative genomics	2016	5.5
BEAST: divergence dating using genomic sequence data, Utrecht University	2016	0.2
The next step in my career	2017	0.6
Introduction to Python for biologists (IPYB05), Glasgow, UK	2018	1.5
Advanced Python for biologists (APYB02), Glasgow, UK	2018	1.5
Illustrator Workshop	2019	0.1
B) SEMINARS & LECTURES		
RIHS seminar How to get the most out of your PhD part 1: industry	2015	0.1
RIHS seminar How to get the most out of your PhD part 2: a career in academia	2015	0.1
C) SYMPOSIA & CONGRESSES		
Science day Internal Medicine	2014	0.25
RIHS PhD Retreat (poster)	2014	0.5
Radboud Science Day (e-poster)	2015	0.25
Center for Molecular and Biomolecular Informatics spring conference (poster)	2015	0.2
KNAW meeting PhD Students on Science 2.0	2015	0.2
Center for Molecular and Biomolecular Informatics autumn conference (oral)	2015	0.2
RIHS PhD Retreat	2015	0.5
TANDEM progress meeting, Craiova, Romania (oral)	2016	1.0
European Molecular Biology Organization Conference Tuberculosis 2016, Paris, France (poster)	2016	1.25
Center for Molecular and Biomolecular Informatics spring conference	2016	0.1
Radboud Center for Infectious Diseases Science Day (e-poster)	2016	0.25

	YEAR(S)	ECTS
TRAINING ACTIVITIES		
C) SYMPOSIA & CONGRESSES		
RIHS PhD Retreat	2016	0.2
TBNET symposium on drug-resistant tuberculosis in Europe		
Vereniging voor Epidemiologie symposium 'Go your own way'	2016	0.5
International Conference on the Pathogenesis of Mycobacterial	2017	0.25
Infections, Stockholm, Sweden (poster)	2017	0.1
RIHS PhD Retreat	2017	1.0
European Congress of Clinical Microbiology and Infectious	2017	0.5
Diseases, Madrid, Spain (e-poster & poster)	2018	1.0
Center for Molecular and Biomolecular Informatics spring conference (oral)	2018	0.2
Union World Conference on Lung Health, The Hague, Netherlands (oral)	2018	1.0
Symposium 'Radboud New Frontiers 2018, Big data, better healthcare?'	2018	0.25
Science day Infectious Diseases & Global Health (oral)	2018	0.25
D) OTHER		
Journal Club Center for Molecular and Biomolecular Informatics	2014–2018	1.5
Journal Club Health Evidence ('junior refereren')	2014–2017	6.0
Center for Molecular and Biomolecular Informatics comparative genomics lunch meetings	2014-2018	6.5
Cytokine meetings Internal Medicine	2014-2017	2.0
'Broodje mycobacteriën' (tuberculosis / nontuberculous mycobacteria lunch meeting)	2016-2018	0.5
TEACHING ACTIVITIES		
E) LECTURING		
BSc course Global Health MED-MIN06	2014,2016	0.2
BSc course Infection and Host Defense 5DT05	2015	0.1
F) SUPERVISION OF INTERNSHIPS / OTHER		
Project supervisor BMS Minor Clinical Research, Radboud University Hilde van Ras, Bo van Santvoort en Marloes Taken	2015–2016	1.0
Project supervisor BMS Minor Clinical Research, Radboud University Monse Wieland, Lieke Vissers, Mariken de Wit en Francine van Wifferen	2016–2017	1.0
OTHER		
RIHS PhD Council member	2015-2017	2.5
Co-organizing a 2-day PhD Retreat	2015-2017	3
College Tour-like interview with Louise Gunning at RIHS PhD Retreat	2016	0.3
College Tour-like interview with Ronald Plasterk at RIHS PhD Retreat	2017	0.3
TOTAL		53.15

Research data management

The clinical data obtained during my PhD at the Radboud university medical center have been captured and stored in REDCap (Research Electronic Data Capture), a secure web application for building and managing online surveys and databases, which is managed by colleagues at the St George's University of London. The privacy of the participants in this study is warranted by use of encrypted and unique individual subject codes. Raw sequencing data described in Chapter 2, 3 and 5 are available through the Sequencing Read Archive (SRA) and are accessible via the sample numbers described in these chapters. The raw sequencing data used for Chapters 4 and 6 are stored on a server belonging to the Center for Molecular and Biomolecular Informatics (CMBI) at the Radboud university medical center, and will be made publically available through the SRA or the European Nucleotide Archive after publication of the respective manuscripts. We are also in the process of archiving these data in the new secured Digital Research Environment (DRE) of the Radboud university medical center. The scripts used to analyse the data are stored on Evernote, application software designed for note taking, organizing, task lists, and archiving. These will also be stored in the DRE. Published data generated or analyzed in this thesis are part of published articles and their additional files and scripts are available from the associated corresponding authors upon reasonable request.

List of publications

Ruth MM, van Rossum M, Koeken VACM, Pennings L, Svensson EM, **Ruesen C**, Bowles EC, Wertheim HFL, Hoefsloot W, van Ingen J. Auranofin activity exposes thioredoxin reductase as a viable drug target in *Mycobacterium abscessus*. *Antimicrob Agents Chemother*. 2019 Jul 1.

Zweijpfenning SMH, Schildkraut JA, Coolen JPM, **Ruesen C**, Koenraad E, Janssen A, Ruth MM, de Jong AS, Kuipers S, Aarnoutse RE, Magis-Escurra C, Hoefsloot W, van Ingen J. Failure with acquired resistance of an optimised bedaquiline-based treatment regimen for pulmonary *Mycobacterium avium* complex disease. *Eur Respir J*. 2019 Apr 18.

Ruesen C*, Chaidir L*, Dutilh BE, Ganiem AR, Andryani A, Apriani L, Huynen MA, Ruslami R, Hill PC, van Crevel R, Alisjahbana B. Use of whole-genome sequencing to predict *Mycobacterium tuberculosis* drug resistance in Indonesia. *J Glob Antimicrob Resist*. 2019 Mar;16:170-177.

Ruesen C*, Riza AL*, Florescu A, Chaidir L, Editoiu C, Aalders N, Nicolosu D, Grecu V, Ioana M, van Crevel R, van Ingen J. Linking minimum inhibitory concentrations to whole genome sequence-predicted drug resistance in *Mycobacterium tuberculosis* strains from Romania. *Sci Rep*. 2018 Jun 26;8(1):9676.

Ruesen C, Chaidir L, van Laarhoven A, Dian S, Ganiem AR, Nebenzahl-Guimaraes H, Huynen MA, Alisjahbana B, Dutilh BE, van Crevel R. Large-scale genomic analysis shows association between homoplastic genetic variation in *Mycobacterium tuberculosis* genes and meningeal or pulmonary tuberculosis. *BMC Genomics*. 2018 Feb 5;19(1):122.

Van Laarhoven A, Dian S, Aguirre-Gamboa R, Avila-Pacheco J, Ricaño-Ponce I, **Ruesen C**, Annisa J, Koeken VACM, Chaidir L, Li Y, Achmad TH, Joosten LAB, Notebaart RA, Ruslami R, Netea MG, Verbeek MM, Alisjahbana B, Kumar V, Clish CB, Ganiem AR, van Crevel R. Cerebral tryptophan metabolism and outcome of tuberculous meningitis: an observational cohort study. *Lancet Infect Dis*. 2018 May;18(5):526-535.

Van Laarhoven A*, Dian S*, **Ruesen C**, Hayati E, Damen MSMA, Annisa J, Chaidir L, Ruslami R, Achmad TH, Netea MG, Alisjahbana B, Rizal Ganiem A, van Crevel R. Clinical Parameters, Routine Inflammatory Markers, and LTA4H Genotype as Predictors of Mortality Among 608 Patients With Tuberculous Meningitis in Indonesia. *J Infect Dis*. 2017 Apr 1;215(7):1029-1039.

Te Brake L*, Dian S*, Ganiem AR, **Ruesen C**, Burger D, Donders R, Ruslami R, van Crevel R, Aarnoutse R. Pharmacokinetic/pharmacodynamic analysis of an intensified regimen containing rifampicin and moxifloxacin for tuberculous meningitis. *Int J Antimicrob Agents*. 2015 May;45(5):496-503.

Delsing CE, **Ruesen C**, Boeree MJ, van Damme PA, Kuipers S, van Crevel R. An African woman with pulmonary cavities: TB or not TB? *Neth J Med*. 2014 Oct;72(8):426-8.

Verheyden G, **Ruesen C**, Gorissen M, Brumby V, Moran R, Burnett M, Ashburn A. Postural alignment is altered in people with chronic stroke and related to motor and functional performance. *J Neurol Phys Ther*. 2014 Oct;38(4):239-45.

Ruesen C, van Gageldonk-Lafeber AB, de Vries G, Erkens CG, van Rest J, Korthals Altes H, de Neeling H, Kamst M, van Soolingen D. Extent and origin of resistance to anti-tuberculosis drugs in the Netherlands, 1993 to 2011. *Euro Surveill*. 2014 Mar 20;19(11).

Ruesen C*, Burhan E*, Ruslami R, Ginanjar A, Mangunnegoro H, Ascobat P, Donders R, van Crevel R, Aarnoutse R. Isoniazid, rifampin, and pyrazinamide plasma concentrations in relation to treatment response in Indonesian pulmonary tuberculosis patients. *Antimicrob Agents Chemother*. 2013 Aug;57(8):3614-9.

Correspondence

Aarnoutse R, **Ruesen C**, Burhan E, van Crevel R, Ruslami R. Reply to "strategy to limit sampling of antituberculosis drugs instead of determining concentrations at two hours postingestion in relation to treatment response". *Antimicrob Agents Chemother*. 2014;58(1):629-30.

* Shared first authorship

Institute for Health Sciences
Radboudumc